A storytelling machine?

Xavier Bost

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 536 « Sciences et Agrosciences »
Laboratoire Informatique d'Avignon (EA 4128)

A storytelling machine?

Automatic video summarization: the case of TV series

par

Xavier BOST

Soutenue publiquement le 23 novembre 2016 devant un jury composé de :

M.	Guillaume GRAVIER	Directeur de Recherche, IRISA, Rennes	Rapporteur (Président)
M.	Éric GAUSSIER	Professeur, LIG, Grenoble	Rapporteur
M^{me}	Martha LARSON	Senior Researcher, MMC, Delft, Pays-Bas	Examinateur
M.	Hervé BREDIN	Chargé de Recherche, LIMSI, Orsay	Examinateur
M.	Vincent LABATUT	Maître de Conférences, LIA, Avignon	Examinateur
M.	Georges LINARÈS	Professeur, LIA, Avignon	Directeur de thèse
M.	Damien MALINAS	Maître de Conférences, CNE, Avignon	Encadrant
M.	Serigne GUEYE	Maître de Conférences, LIA, Avignon	Encadrant

Laboratoire Informatique d'Avignon

Contents

Remerciements

Je tiens tout d'abord à remercier l'ensemble des membres de mon jury de thèse, et en particulier les rapporteurs, pour leurs remarques détaillées et constructives.

Mes remerciements vont ensuite à mon équipe d'encadrement : Georges en premier lieu, qui m'a proposé un problème de recherche original et ouvert, en adéquation avec mon profil pluridisciplinaire, et qui a su trouver la bonne distance pour m'accompagner pendant ces trois années ; Serigne ensuite, pour sa disponibilité sans faille ; et Damien, pour la nécessaire ouverture aux importantes questions soulevées par la sociologie des publics de la culture. Un remerciement tout particulier enfin à Vincent, qui a rejoint l'aventure en route, et a su apporter à ce travail son expertise en réseaux complexes.

Je remercie également Martha, et toute l'équipe du *Multimedia Computing Group* de l'Université Technologique de Delft, qui m'ont réservé un accueil chaleureux et bénéfique lors de mon séjour de recherche aux Pays-Bas.

Une pensée particulière également pour ceux qui m'ont accompagné lors de mes premiers contacts avec la recherche scientifique : Marc El-Bèze et Renato de Mori, qui m'ont beaucoup appris lors de mon stage de Master ; Fabrice Lefèvre, auprès duquel j'ai pris un tout premier contact avec des questions scientifiques ouvertes dans le cadre de mon projet de Master.

Je remercie également Laura, Mathilde et Danny, les trois étudiants du Master *Publics de la culture et communication* qui ont pris une part active à l'élaboration et à la diffusion du questionnaire sur lequel repose l'essentiel de l'apport expérimental de ce travail.

D'un point de vue plus institutionnel, mes remerciements vont à la Fédération de Recherche *Agorantic* de l'Université d'Avignon, qui a financé ce travail, et au *Laboratoire Informatique d'Avignon*, pour m'avoir offert des conditions de travail très favorables, notamment grâce à l'efficacité de son équipe administrative.

Une pensée enfin pour ma famille, qui a su m'accompagner avec bienveillance dans cette aventure qu'a été ma reprise d'études dans un domaine très éloigné de ma formation initiale : d'abord entreprise par curiosité intellectuelle, cette seconde formation a pris en Doctorat une véritable consistance professionnelle, propre à m'ouvrir de nouvelles et stimulantes perspectives.

Personal Bibliography

International Journals

- *Remembering Winter Was Coming: Character-oriented Video Summaries of TV Series*

  Bost Xavier, Gueye Serigne, Labatut Vincent, Larson Martha, Linarès Georges, Malinas Damien, Roth Raphaël

  IEEE Transactions on Multimedia – [Being submitted]

  2016

- *Multiple topic identification in human/human conversations*

  Bost Xavier, Senay Grégory, El-Bèze Marc, De Mori Renato

  Computer, Speech and Language

  2015

International Conferences / Workshops

- *Narrative Smoothing: Dynamic Conversational Network for the Analysis of TV Series Plots*

  Bost Xavier, Labatut Vincent, Gueye Serigne, Linarès Georges

  International Workshop on Dynamics in Networks

  DyNo 2016, in conjunction with the

  2016 IEEE/ACM International Conference **ASONAM**

  August 18–21, 2016, San Francisco, USA

- *Audiovisual speaker diarization of TV series*

  Bost Xavier, Linarès Georges, Gueye Serigne

  IEEE International Conference on Audio, Speech and Signal Processing

  ICASSP 2015

  April 19–24, 2015, Brisbane, Australia

- *Constrained speaker diarization of TV series based on visual patterns*

  Bost Xavier, Linarès Georges

  IEEE/ISCA Speech and Language Technology Workshop

  SLT 2014

  December 7–10, 2014, South Lake Tahoe, USA

- *Multiple topic identification in telephone conversations*

  Bost Xavier, El-Bèze Marc, De Mori Renato

  International Confererence of the Speech Communication Association, ISCA

  InterSpeech 2013

  August 25–29, 2013, Lyon, France

National Conferences / Workshops

- *Détection de locuteurs dans les séries TV*

  Bost Xavier, Linarès Georges

  COnférence en Recherche d'Information et Applications

  CORIA 2015

  March 18–20, 2015, Paris, France

- *Catégorisation multi-thématique de dialogues téléphoniques*

  Bost Xavier, El-Bèze Marc, De Mori Renato

  Journées d'Études de la Parole

  JEP 2014

  June 23–27, 2014, Le Mans, France

Challenge submissions

- *LIA@RepLab 2013*

  Cossu Jean-Valère, Bigot Benjamin, Morchid Mohamed, Bost Xavier, Senay Gregory, Bouvier Vincent, Dufour Richard, Torres Juan-Manuel, El-Bèze Marc

  RepLab Overview

  CLEF 2013

  September 23–26, 2013, Valencia, España

Résumé

Ces dix dernières années, les séries télévisées sont devenues de plus en plus populaires. Par opposition aux séries TV classiques composées d'épisodes autosuffisants d'un point de vue narratif, les séries TV modernes développent des intrigues continues sur des dizaines d'épisodes successifs. Cependant, la continuité narrative des séries TV modernes entre directement en conflit avec les conditions usuelles de visionnage : en raison des technologies modernes de visionnage, les nouvelles saisons des séries TV sont regardées sur de courtes périodes de temps. Par conséquent, les spectateurs sur le point de visionner de nouvelles saisons sont largement désengagés de l'intrigue, à la fois d'un point de vue cognitif et affectif. Une telle situation fournit au résumé de vidéos des scénarios d'utilisation remarquablement réalistes, que nous détaillons dans le Chapitre 1. De plus, le résumé automatique de films, longtemps limité à la génération de bande-annonces à partir de descripteurs de bas niveau, trouve dans les séries TV une occasion inédite d'aborder dans des conditions bien définies ce qu'on appelle le *fossé sémantique* : le résumé de médias narratifs exige des approches orientées contenu, capables de jeter un pont entre des descripteurs de bas niveau et le niveau humain de compréhension. Nous passons en revue dans le Chapitre 2 les deux principales approches adoptées jusqu'ici pour aborder le problème du résumé automatique de films de fiction. Le Chapitre 3 est consacré aux différentes sous-tâches requises pour construire les représentations intermédiaires sur lesquelles repose notre système de génération de résumés : la Section 3.2 se concentre sur la segmentation de vidéos, tandis que le reste du chapitre est consacré à l'extraction de descripteurs de niveau intermédiaire, soit orientés saillance (échelle des plans, musique de fond), soit en relation avec le contenu (locuteurs). Dans le Chapitre 4, nous utilisons *l'analyse des réseaux sociaux* comme une manière possible de modéliser l'intrigue des séries TV modernes : la dynamique narrative peut être adéquatement capturée par l'évolution dans le temps du réseau des personnages en interaction. Cependant, nous devons faire face ici au caractère séquentiel de la narration lorsque nous prenons des vues instantanées de l'état des relations entre personnages. Nous montrons que les approches classiques par fenêtrage temporel ne peuvent pas traiter convenablement ce cas, et nous détaillons notre propre méthode pour extraire des réseaux sociaux dynamiques dans les médias narratifs. Le Chapitre 5 est consacré à la génération finale de résumés orientés personnages, capables à la fois de refléter la dynamique de l'intrigue et de ré-engager émotionnellement les spectateurs dans la narration. Nous évaluons notre système en menant à une large échelle et dans des conditions réalistes une enquête auprès d'utilisateurs.

Mots-clés : *résumé de vidéos*, *séries* TV, *analyse de l'intrigue*, *analyse des réseaux sociaux*, *segmentation en locuteurs*.

Abstract

THESE past ten years, TV series became increasingly popular. In contrast to classical TV series consisting of narratively self-sufficient episodes, modern TV series develop continuous plots over dozens of successive episodes. However, the narrative continuity of modern TV series directly conflicts with the usual viewing conditions: due to modern viewing technologies, the new seasons of TV series are being watched over short periods of time. As a result, viewers are largely disengaged from the plot, both cognitively and emotionally, when about to watch new seasons. Such a situation provides video summarization with remarkably realistic use-case scenarios, that we detail in Chapter 1. Furthermore, automatic movie summarization, long restricted to trailer generation based on low-level features, finds with TV series a unprecedented opportunity to address in well-defined conditions the so-called *semantic gap*: summarization of narrative media requires content-oriented approaches capable to bridge the gap between low-level features and human understanding. We review in Chapter 2 the two main approaches adopted so far to address automatic movie summarization. Chapter 3 is dedicated to the various subtasks needed to build the intermediary representations on which our summarization framework relies: Section 3.2 focuses on video segmentation, whereas the rest of Chapter 3 is dedicated to the extraction of different mid-level features, either saliency-oriented (shot size, background music), or content-related (speakers). In Chapter 4, we make use of *social network analysis* as a possible way to model the plot of modern TV series: the narrative dynamics can be properly captured by the evolution over time of the social network of interacting characters. Nonetheless, we have to address here the sequential nature of the narrative when taking instantaneous views of the state of the relationships between the characters. We show that standard time-windowing approaches can not properly handle this case, and we detail our own method for extracting dynamic social networks from narrative media. Chapter 5 is dedicated to the final generation and evaluation of character-oriented summaries, both able to reflect the plot dynamics and to emotionally re-engage viewers into the narrative. We evaluate our framework by performing a large-scale user study in realistic conditions.

Keywords: *video summarization*, TV *series*, *plot analysis*, *social network analysis*, *speaker diarization*.

Chapter 1

Introduction:
Remembering winter was coming

Contents

According to the *Oxford Dictionary*, a summary is "a brief statement or account of the main points of something"[1]. Despite its concision, such a definition mentions every essential feature of a summary:

- **Relativity**

  A summary is by definition a summary *of* something, which we will denote as the *source* it is artificially built on. The source may be conveyed by various media streams: textual, oral, or even video. The summary may consist in a full reformulation of its source, and be *abstractive*, or just contain excerpts, possibly re-arranged following some editing rules. The latter type is commonly denoted as *extractive* summary. In the special case of a video source, the excerpts included in the extractive summary can be still or moving images: the resulting summaries are sometimes denoted, respectively, as *keyframes* and *video skims*. In case of abstractive summarization, the medium of the summary may be different from the source medium: movies for instance may be verbally summarized.

- **Brevity**

  Secondly, a summary has to be *brief*, at least shorter than its source. Such a property suggests that in some of their uses, summaries respond to time constraints.

- **Capturing "important parts"**

  Finally, a summary is expected to report the *main points* contained in its source.

This last property is by far the most subjective one: summaries have to convey the most important parts of their source, but such a notion of importance is by nature relative to human assessment and dependent on specific use-case scenarios.

For instance, movie trailers can be considered as extractive summaries: they are built upon the movie rushes; they last only a few minutes and provide us with engaging sequences. Nonetheless, trailers are designed to make people want to watch a film they haven't viewed yet: for such a promotional purpose, trailers avoid unveiling the full development of the plot, and especially how it ends; instead, they will focus, often without any respect for the narrative chronology, on engaging scenes, sometimes resulting in a distorted overview of the movie mood. If people were to view an extractive summary of a movie they have already watched, for instance in order to remember the plot, the purpose of the summary would be quite different, along with the notion of "important scenes", that would for sure, unlike trailers, include the final resolution scene.

The content of a summary is therefore highly dependent on its expected purpose and on the considered use-case scenarios, either implicitly or explicitly formulated. From a computational perspective, this point is of primary importance: on the one hand, the role assigned to the summary directly impacts the choice of features considered when modeling the source; on the other hand, the evaluation of the automatically generated summaries is highly concerned with the purpose they are expected

[1] http://www.oxforddictionaries.com/definition/english/summary

to achieve. According to (Truong et Venkatesh, 2007) in their review of video abstraction techniques, “the main focus of the evaluation process should be application-dependent”.

From the perspective of automatic summarization, modern TV series provide us with remarkably realistic, though quite challenging, use-case scenarios. In this introduction, we motivate the need for video summaries of TV series from the user’s perspective, before briefly describing the choices we made in this work to automate as much as possible the generation of such summaries.

1.1 The “serial paradox”

These past ten years, TV series became increasingly popular. One week after being broadcast on HBO, the first episode of the fifth season of *Game of Thrones* was (illegally) downloaded 32 million times; TV series now have their own dedicated wikis[2], forums and *YouTube* channels on the Web, along with their own festivals[3]. Moreover, long despised as a minor genre, TV series have recently received critical acclaim, as a unique space of creativity, able to attract even renowned full-length movie directors, such as Jane Campion or Bryan Singer. As shown on Fig. 1.1, for more than half of the people we polled[4], watching TV series is a daily occupation, and more than 80% watch TV series at least once a week.

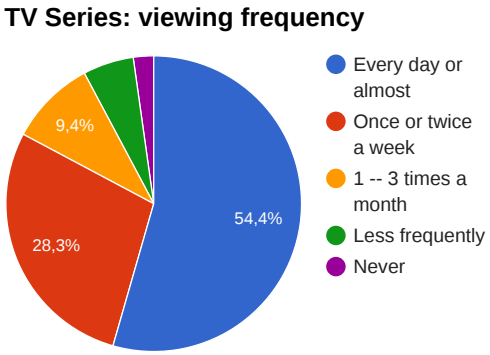


Figure 1.1 *– In the last 12 months, how often have you watched TV series?*

Such a success is probably in part closely related to the cultural changes induced by modern technologies (Malinas, 2015): the extension of high-speed internet connections

[2]See for instance http://gameofthrones.wikia.com/wiki/Game_of_Thrones_Wiki

[3]In France, http://series-mania.fr

[4]All the statistics reported in this Section come from the large-scale (187 participants, mostly students) user study we performed for validation purpose: see Appendix A for a full description of the study, and (Combes, 2013) for a comprehensive sociological study of the TV series audience.

led to unprecedented viewing opportunities. As can be seen from Fig. 1.2, streaming or downloading services are by far the most widely used channels for accessing TV series, and television, in a kind of paradox, is no longer used as the main way of viewing "TV" series.

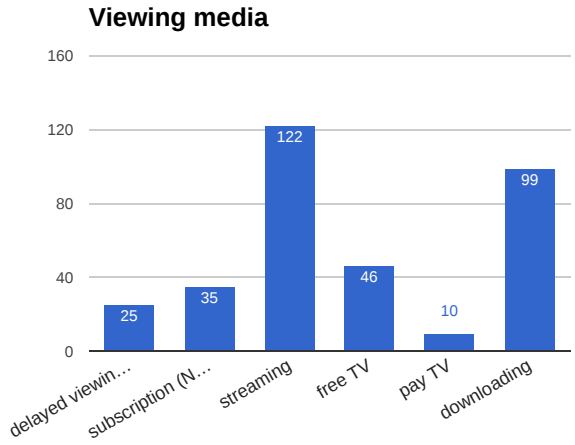


***Figure 1.2** – How do you usually watch TV series? (multiple answers allowed)*

Unlike television, streaming and downloading services give control to the user, not only over the contents he wants to watch, but also over the viewing frequency. The typical dozen of episodes that a TV series season contains is usually being watched over a much shorter period of time than the usual two months it is being aired on television.

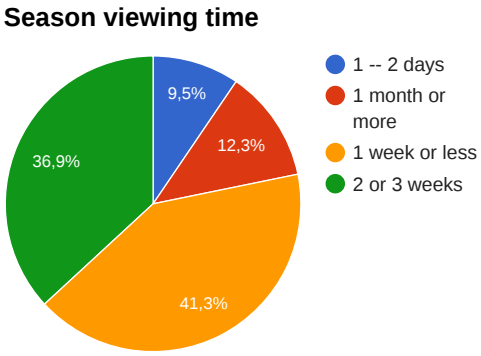


***Figure 1.3** – In case of delayed viewing, how long does it take you on average to view a full season?*

As shown on Fig. 1.3, for 41% of the people we polled, a whole season (about 10 hours of viewing in average) is watched in only one week, with 2-3 successive episodes at once (not reported on Fig. 1.3), and for 9% of them, the viewing period of a season is even shorter (1-2 days), resulting in the so-called "binge-watching" phenomenon. In

summary, television is no longer the main channel used to watch TV series, resulting in short viewing periods of the new seasons of TV series, usually released once a year.

TV series come in four main types:

- **Classical TV series** consist of narratively self-contained episodes with recurring characters (*e. g. The Avengers*, *Columbo*, *Mission: Impossible*).
- **Anthologies** consist of narratively self-contained units with no recurring characters, either episodes (*e. g. The Twilight Zone*), or seasons (*e. g. American Horror Story*, *True detective*, *Fargo*). The latter case corresponds to the so-called *modern anthology*.
- **Serials** rely on a continuous story, each episode contributing to the plot (*e. g. Breaking bad*, *Game of Thrones*, *House of Cards*).
- The last one is a **mix** between the classical and serial genres: each episode develops its own plot, but also contributes to a secondary, continuous story (*e. g. House M. D.*, *Person of Interest*).

Whereas classical TV series, with their standalone episodes, are still represented nowadays, they are by far not as popular as TV serials, based on continuous, possibly complex stories usually spanning several episodes, when not several seasons. As can be seen from Fig. 1.4, 66% of the people we polled prefer TV serials to series with standalone episodes.

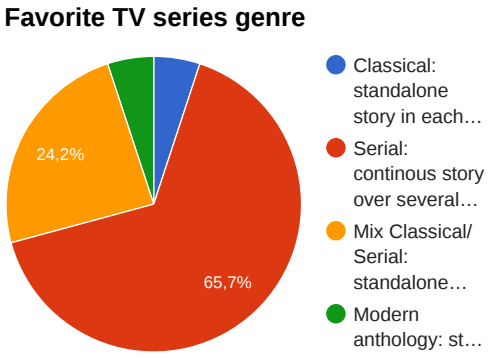


***Figure 1.4** – What is your favorite TV series genre?*

Yet, the narrative continuity of TV serials directly conflicts the usual viewing conditions described above, resulting in what we call the “serial paradox”: *highly continuous from a narrative point of view,* TV *serials, like any other* TV *series, are typically watched in quite a discontinuous way*; when the episodes of a new season are released, the time elapsed since viewing the last episode usually amounts to several months, when not nearly one year.

1.2 The cold-start phenomenon

As a first major consequence of this serial paradox, viewers are likely to have forgotten to some extent the plot of TV serials when they are, at last, about to know what comes next: as shown on Fig. 1.5, nearly 60% of the people we polled feel the need to remember the main events of the plot before viewing the new season of a TV serial.

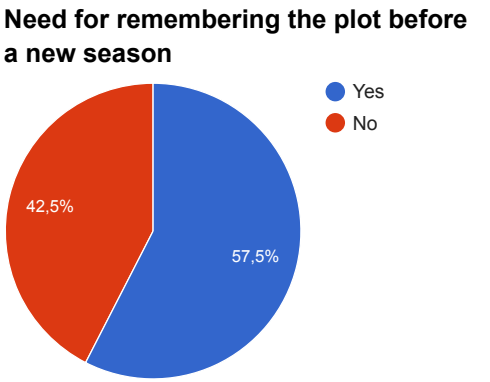


Figure 1.5 *– Before viewing the new season of a TV serial, do you feel the need to remember the plot of the previous seasons?*

Fig. 1.6 shows the main information channels viewers consult to address such a cognitive disengagement.

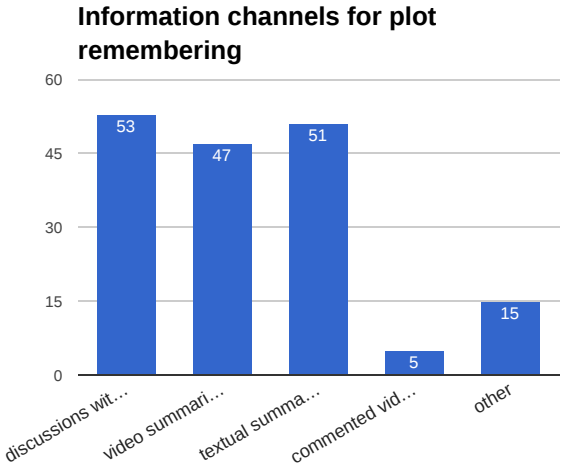


Figure 1.6 *– Before viewing the new season of a TV serial, which information channel(s) do you use for remembering the plot of the previous seasons? (multiple answers allowed)*

Whereas discussing with friends is a common practice to help to remember the plot of the previous seasons (49% of the polled people), the recaps available online are also

extensively used to fill such a need: before viewing a new season, about 48% of the people read textual synopsis, mainly in *Wikipedia* (not reported on Fig. 1.6), and 43% of the people watch video recaps, either "official" or handmade, often on *YouTube* (not reported on Fig. 1.6). Interestingly, none of these ways of reducing the "cognitive loading" that the new season could induce excludes the others, and people commonly use multiple channels of information to remember the plot of TV serials.

Furthermore, the time elapsed since watching the previous season may be so long that the desire to watch the next one gets weaker, possibly resulting in a disaffection for the whole TV serial. The website *GraphTV*[5] allows to visualize the average ratings on the *Internet Movie DataBase* (*IMDb*[6]) of TV series episodes, along with season trendlines. Fig. 1.7, 1.8 and 1.9 respectively show the plots of such average ratings for the three TV serials considered in this work, namely *Breaking Bad*, *Game of Thrones* and *House of Cards*, with different colors for successive seasons. As can be seen from all three figures, the average rating of each episode surprisingly tends to look like a linear function of its rank in the season, and seems partially independent of the episode's intrinsic qualities.

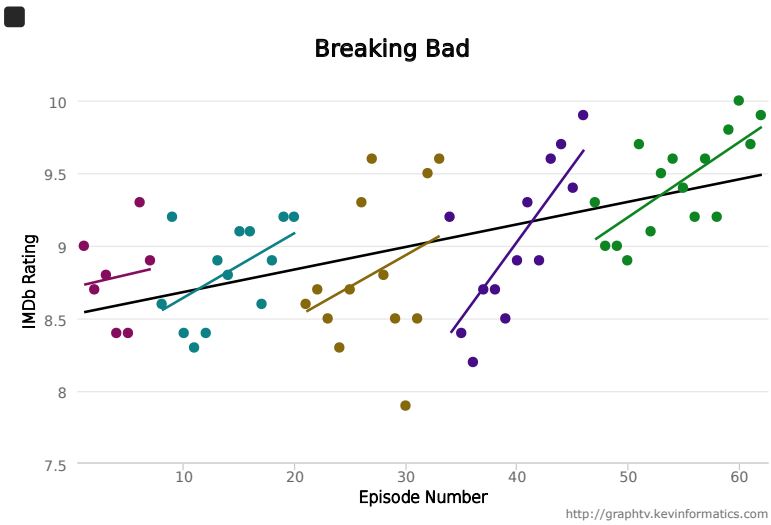


***Figure 1.7** – Average ratings of* Breaking Bad *episodes on IMDb along with global and season trendlines.*

A first possible explanation of such a phenomenon would be that only those people who liked the very first episodes keep on watching, and rating, the following ones, resulting both in fewer and higher ratings over time. Nevertheless, such an explanation, known as the "survivor bias"[7], would only stand for the very first episodes of the first season. After these first episodes, the audience is expected to stabilize, and the *IMDb* ratings shown on *GraphTV* actually rely on a remarkably stable number of votes.

Instead, the increasing season trendlines of the average ratings can also be inter-

[5] http://graphtv.kevinformatics.com

[6] http://www.imdb.com. Ratings range from 1 to 10.

[7] http://www.spoilertv.com/2014/09/do-tv-series-get-better-or-worse-over.html

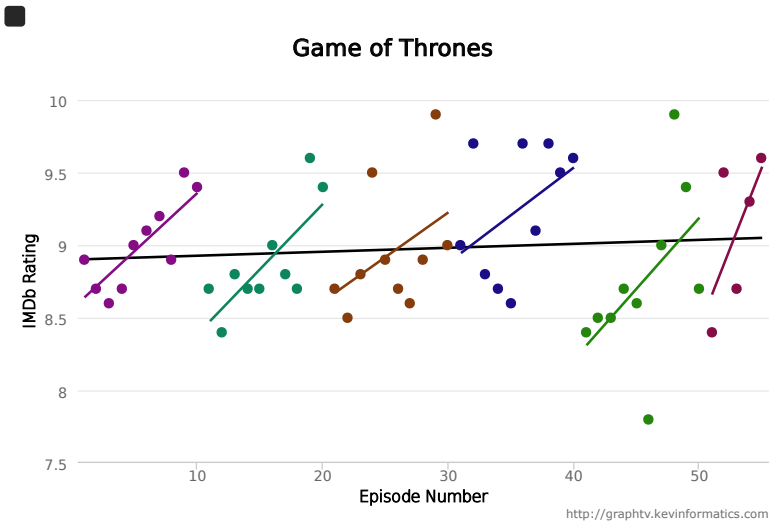


***Figure 1.8** – Average ratings of* Game of Thrones *episodes on IMDb along with global and season trendlines.*

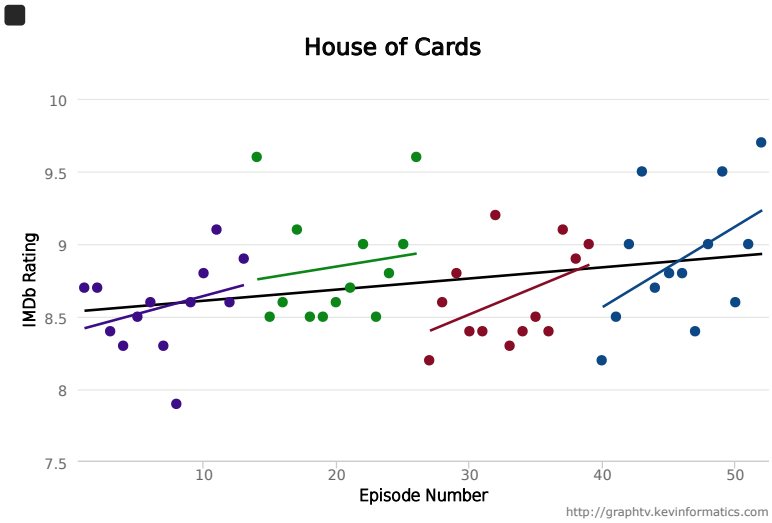


***Figure 1.9** – Average ratings of* House of Cards *episodes on IMDb along with global and season trendlines.*

preted as exhibiting a kind of "cold-start" phenomenon. Part of the success of TV serials rely on their "hooking" effect[8] while viewing a season. Conversely, the viewer's engagement in the series is expected to radically drop off between seasons. Despite intensive advertising campaigns around the new season and possible use of cliffhangers during the previous season finale, the audience needs to get re-immersed in the series' universe and storylines. According to Fabrice Gobert, the director of the French TV serial *Les revenants*, "Writing the first episode of the first season is not an easy thing, but

[8] http://www.prnewswire.com/news-releases/do-you-know-when-you-were-hooked-netflix-does-300147700.html

the first episode of the second season is not easy either, because we have to make the spectator want to re-immerse himself in the series."[9]

To summarize, the unavoidable waiting time between successive seasons of TV serials results in radical drop-off in viewers' engagement, both cognitive and affective, and draw two major use-cases for summaries of TV serials, that should act both as content-covering informative synopses and emotionally engaging promotional trailers.

1.3 Automatic summaries for viewer's re-engagement

In this work, we investigate ways of automatically generating video summaries that can effectively address viewers' drop-off between seasons of TV serials.

The reminding effect expected from our summaries first requires the plot of TV serials to be effectively modeled and covered, but for viewers who were at some point familiar with the narrative content. In such a use-case, the generated summaries are only expected to act as effective plot reminders, and not to dispense viewers from watching the original source, for instance to catch up with missed episodes: as shown on Fig. 1.10, only 12% of the people we polled use the video summaries available online in this way. As stated above, modern technologies tend to relieve the users from the official broadcasting, and it remains still possible to fully watch missed episodes. In such a use-case scenario, the plot modeling step is definitely not as challenging as if it were intended to support the building of substituting summaries, semantically self-contained and where no past acquaintance with the story could be assumed.

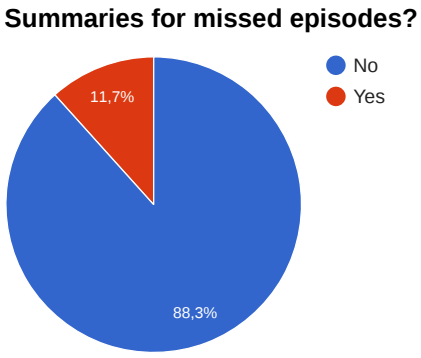


Figure 1.10 – *Do you sometimes watch summaries to catch up with missed episodes?*

[9] Radio interview on *France Culture*, available on http://www.franceculture.fr/emissions/ping-pong/femme-fatale-et-revenants-avec-diane-kruger-fabrice-gobert

Our first contribution is the use of *social network analysis* for such a plot modeling purpose. The narrative dynamics of TV serials relies on evolving social configurations of interacting characters. Accordingly, we base in Chapter 4 the analysis of the plot on the dynamic conversational network of characters, as built upon the verbal interactions between speakers. Such a way of modeling the plot requires first to automatically detect the speakers involved. We perform speaker detection in a fully unsupervised way, as a speaker diarization task in a multimedia context: our second contribution is the use in Section 3.5 of the multiple modalities available in the multimedia stream to automatically perform the speaker diarization subtask.

Nonetheless, the trailer effect expected from our summaries can not only rely on plot coverage, but has to be supported by expressive, emotionally engaging, video excerpts. For this purpose, we make use of two additional features based on film grammar: shot size and background music, both described in Chapter 3, respectively in Sections 3.3 and 3.4. Such features are expected to highlight engaging sequences, able to support the "trailer" effect also expected from fully revival summaries.

Our last contribution is the large-scale user study (Chapter 5) we performed in a real-case scenario to evaluate the summaries of TV series we generate. The evaluation of our framework in realistic conditions led to well-supported conclusions, potentially useful for future work.

On the whole, even though we make use in this work of some image processing techniques, the audio source turns out to be valuable for TV serial summarization: as stated above, the way we model the narrative relies on a dynamic network on interacting speakers and could also, as suggested in Chapter 6, benefit from the analysis of the speech content. Furthermore, emotionally engaging sequences are usually emphasized in movies by background music and can additionally be detected from the low-level audio features we detail in Chapter 2.

1.4 Corpus

Very few corpora of annotated TV series episodes are available. Significantly, many of them focus on classical TV series with narratively self-contained episodes: for instance, (Tapaswi et al., 2012) and (Roy et al., 2014) focus on the popular sitcom *The Big Bang Theory*; the experimental part of (Ercolessi et al., 2011) relies on a few episodes of the sitcom *Ally McBeal* with annotations made publicly available. However, (Sivic et al., 2009) and (Ferrari et al., 2009) focus on a few episodes of *Buffy the Vampire Slayer*, a TV series that combines the classical and serial genres. Some episodes of the popular TV serial *Game of Thrones* have been annotated in a linguistic perspective described in (Roy et al., 2014), and (Cour et al., 2008) provide visual annotations for the two TV serials *Lost* and *Alias*.

We choose to focus on three popular American TV serials:

- *Breaking Bad* (denoted hereafter BB): 60 episodes (5 seasons) released.

- *Game of Thrones* (GoT): 60 episodes (6 seasons) released so far.
- *House of Cards* (HoC): 42 episodes (4 seasons) released so far.

Besides their popularity, the first reason for specifically focusing on these three TV series is their obvious narrative continuity between episodes and even seasons. Unlike classical and mixed classical/serial TV series, none of them contains episodes that can be regarded as self-sufficient from a narrative point of view.

Secondly, they respectively belong to very different genres: BB is categorized in *Wikipedia* both as a *crime drama*, a *contemporary western* and a *black comedy*[10]; GoT is a *fantasy drama*[11] and HoC a *political drama*[12]. Such a diversity is expected to prevent us from designing algorithms that would have been too specific and difficult to generalize.

Finally, BB is now completed: the last part of the final fifth season was released in 2013. On the opposite, the other two ones, GoT and HoC, are not yet completed at the time we are writing. Indeed, we suspected that the summarization process could be impacted by the possible knowledge of the whole development of the story from the very beginning until the final outcome.

***Table 1.1** – Annotations available in our corpus, either manually (man.) or automatically (auto.) introduced.*

Modality	Task	TV serial					
		BB		GoT		HoC	
		S01 **(7)**	S02–03 **(26)**	S01 **(10)**	S02–05 **(40)**	S01 **(13)**	S02 **(13)**
Image	*shot segmentation*	man.	auto.	man.	auto.	man.	auto.
	shot similarity	man.	auto.	man.	auto.	man.	auto.
	scene segmentation	man.	man.	man.	man.	man.	man.
	face detection	auto.	auto.	auto.	auto.	auto.	auto.
Text	*subtitle content*	auto.	auto.	auto.	auto.	auto.	auto.
Audio	*speaker diarization*[a]	**man.**	man.	**man.**	man.	**man.**	man.
	verbal interactions[b]	both[c]	both[d]	both[e]	auto.	both[f]	auto.
	music tracking	auto.	auto.	auto.	auto.	auto.	auto.

[a] The recaps inserted at the beginning of the HoC episodes are not annotated for this task.
[b] The recaps inserted at the beginning of the HoC episodes are not annotated for this task.
[c] S01E04, S01E06: man. – others: auto.
[d] S02E03, S02E04: man. – others: auto.
[e] S01E03, S01E07, S01E08: man. – others: auto.
[f] S01E01, S01E07, S01E11: man. – others: auto.

For the annotation purpose, we acquired the DVDs of the five seasons of BB, the first five seasons of GoT and the first two seasons of HoC. Table 1.1 reports the episodes

[10] https://en.wikipedia.org/wiki/Breaking_Bad
[11] https://en.wikipedia.org/wiki/Game_of_Thrones
[12] https://en.wikipedia.org/wiki/House_of_Cards_(U.S._TV_series)

we annotated, either manually (denoted *man.*) or automatically (denoted *auto.*) by using the algorithms we describe in this document. The number of episodes annotated is mentioned in parentheses (4^{th} line). The subtitle content is retrieved by using a standard OCR tool based on the *Tesseract* engine[13]. When we manually labeled each subtitle according to the corresponding speaker, we sometimes adjusted the subtitle boundaries so that to match the audio signal. For the set of episodes concerned, we put the mention *man.* in bold characters. On the whole, we annotated 109 episodes, and we made publicly available online[14] the resulting corpus of annotations.

1.5 Organization of the document

The document is organized as follows.

In Chapter 2, we review some related works. We distinguish between two major ways of addressing the video summarization task, with a focus on movie summarization. The first class of approaches attempts to isolate salient sequences based on relatively low-level features; the second kind of approaches attempts to model the video stream content, and especially the storylines of the narrative when it comes to model the content of full-length movies. In this chapter, we pay special attention to the use-case scenarios considered, together with the related evaluation protocols.

Chapters 3 and 4 are devoted to all the preliminary subtasks needed for generating both content-covering and engaging summaries.

Chapter 3 focuses both on video segmentation techniques and feature extraction. Segmenting the video stream is first needed for isolating the elementary units candidate for later insertion in the summary; furthermore, some of the feature extraction techniques (speaker diarization, interaction estimate) we use assume some pre-segmented sequences of various types. Every kind of video sequence needed for both purposes is described, along with the extraction techniques we used. Chapter 3 then focuses on feature extraction, both style-based (shot size estimate, music tracking) and content-oriented (speaker diarization).

Chapter 4 details the way the plot of TV serials can be modeled from the dynamic network of interacting speakers. We first describe and evaluate the way verbal interactions can be estimated by jointly applying a set of basic heuristics to the speech turns sequence, once labeled. We then introduce a novel algorithm, named *narrative smoothing*, for building the dynamic network of interacting speakers, which properly handles the narrative sequentiality issue.

Chapter 5 focuses on the summarization process, along with the experimental results. We first show how each character's storyline can be automatically segmented at different scales by considering his/her social neighborhood as it is evolving over time. We then describe how the relevance, both content and style-based, of each candidate

[13] https://github.com/tesseract-ocr/tesseract/wiki
[14] https://dx.doi.org/10.6084/m9.figshare.3471839

unit sequence can be computed from the features described in Chapters 3 and 4. We then detail the summarization algorithm we use, which adapts the standard Maximum Margin Relevance (MMR) algorithm by applying a slightly modified heuristics for solving the associated quadratic knapsack problem. We finally conclude this chapter by detailing the evaluation protocol we designed in order to assess our summaries, along with the experimental results we obtained.

Chapter 2

Related works

Contents

In this section, we review some of the works in the automatic video summarization field, with a special focus on movie skimming, where the summary consists of sequences extracted from the video source, usually reassembled in chronological order. Because the notion of "important parts" expected to be captured in any effective summary remains use-case dependent, we pay special attention to the use-case scenarios considered, together with the related features and evaluation protocol. Comprehensive reviews of works related to video summarization in general can be found in (Truong et Venkatesh, 2007) and (Money et Agius, 2008).

2.1 Highlighting summaries based on saliency

We first focus on highlighting summaries[1]: such summaries do no pretend to fully cover the source content; instead, they aim at preserving "interesting or important events in the video" (Truong et Venkatesh, 2007), usually indirectly detected from low or mid-level features expected to trigger viewers' reactions.

2.1.1 Features for saliency

Though classical approaches in text summarization partially rely on formal features for capturing relevant sentences, such as the title/headings status of a sentence and its location in the document (Edmundson, 1969), text summarization can generally not rely in the first place on low-level formal markers for capturing salient sequences: if punctuation signs such as the exclamation mark are of common use in Western literature, character attributes (italic, bold, upper-case...) are not as common when it comes to emphasize some sentences. As a result, textual saliency can generally not be captured from such low-level features, and text summarization is left with directly dealing with the lexical or semantic content of the source.

Instead, videos, because of their multi-modal nature, contain many saliency markers, often computable from relatively low-level features. Such saliency markers provide us with straightforward techniques for building highlighting summaries.

Sport meetings

Before being applied to movies, video highlights were first proposed to summarize video recordings of sport meetings, where salient events can be objectively defined and captured by relying on relatively low-level features.

Some works attempt to directly capture salient events in sport meetings by relying on the specific associated features: for instance, (Assfalg et al., 2003) build a soccer-specific goal detection model by tracking both ball motion and players' position. Similarly, (Chang et al., 2002) make use of HMMs to categorize shots of baseball video recordings into event classes to detect highlights.

[1] Denoted *video previews* in (Tsoneva et al., 2007).

A more generic alternative than such domain-specific models is introduced in (Li et Sezan, 2001) to detect *play* events in various sport recordings: rather than trying to capture such events from their computable content, the authors show that typical sequences of camera alternations correspond to *play* events in many sport recordings. Summaries focusing on these events are then automatically built by applying both rule-based, and HMMs-based approaches to baseball, American football and Japanese sumo wrestling video recordings.

A highly generic framework is introduced in (Hanjalic et Xu, 2005) to automatically build video highlights of sport meetings from the reactions that salient events are likely to trigger. Some of the features used by the authors remain strongly related to the content of salient events: objective motion activity for example target the participants' behavior; but some other features are clearly content-independent: high subjective motion activity (camera-based) and peaks in shot frequency for instance are more related to the way content producers react to salient events by using emphasizing filming techniques. Similarly, high audio energy is not directly part of salient events, but allows to indirectly capture reactions to salient events from the live audience and commentators.

User attention models

Attention models fully achieve the shift from content-related, domain-specific highlighting approaches to more generic techniques, expected to be content-independent and with possible application to movies: such models explicitly aim at providing general-purpose summarization frameworks that dispense to deal with the semantic content of the summary source. Important sequences are expected to be emphasized by formal, low-level features commonly used by content producers to trigger user attention:

1. Visual features:

   Motion, both objective and subjective (camera-based), is commonly used, for instance in (Ma et al., 2002) and (Hanjalic et Xu, 2005), as a relevant feature for capturing engaging sequences likely to trigger human excitement; saliency maps for static shots are used in (Ma et al., 2002) and (Evangelopoulos et al., 2013). (Ma et al., 2002) also consider the size and position of possible faces as correlated with human attention and (Hanjalic et Xu, 2005) include shot frequency in their attention model.

2. Acoustic features:

   Sound energy is the most widely used feature for acoustically capturing salient sequences in the source stream: (Ma et al., 2002), (Hanjalic et Xu, 2005) and (Evangelopoulos et al., 2013) all make use of this feature. The comprehensive attention model detailed in (Ma et al., 2002) relies on some additional acoustic features: the presence of speech and music in the original audio signal is assumed to correspond to salient sequences; music in particular is commonly used by filmmakers "to emphasize the atmosphere of scenes in videos". MFCCs are then extracted for pre-categorizing audio sub-segments into music, speech, or silence.

3. Textual features:

   (Evangelopoulos et al., 2013) consider, in addition to these standard audio-visual features, the textual saliency of movie subtitles as a relevant feature to capture salient sequences: some part-of-speech classes are claimed to be more salient than others. Proper nouns for instance are assumed to be more relevant and salient than stop-words without proper semantic content. Based on word saliency, a textual saliency curve is computed for the whole video stream, before highlights are detected.

The paradigmatic shift from the video content to the viewer's reaction to formally salient sequences is expected to result in more generic summarization frameworks, with possible applications to "automated movie trailer" (Hanjalic et Xu, 2005). Indeed, formal saliency based on low and mid-level features is widely used in movie trailer automatic generation.

Trailers

The features used in trailer automatic generation turn out to be very close to the features targeted by the attention models described above.

For instance, (Chen et al., 2004) automatically generate previews and trailers of action movies from their tempo, defined as a combination of shot frequency, motion activity and audio energy: salient sequences in action movies, relevant for further insertion in the summary, are expected to contain fast-alternating shots with fast motion, sound effects and music. Based on these features, the video stream is first segmented into scenes distributed around action peaks, before each scene is represented in the summary by high-activity shots: the resulting summary is stated to be either a trailer, if all single shots are concatenated, of previews if they are inserted along with the surrounding shots.

(Smeaton et al., 2006) use similar features to discriminate relevant shots for later insertion in action movie trailers: shot frequency and motion intensity are used as relevant visual features for this task; besides, the distribution of each shot over audio classes is used as a relevant acoustic feature: action scenes are expected to contain music rather than speech[2].

Finally, (Irie et al., 2010) make a step further towards automatic generation of trailer-like movie summaries. Shots are first selected from standard low-level visual features based on attention modeling: image color/brightness, motion intensity. Careful attention is then paid to the editing process of the set of selected shots, introduced in the summary so as to produce the most impressive effect rather than in chronological order; furthermore, the theme music, along with the movie title and impressive sounds based on audio signal pitch, energy and MFCCs, are added afterwards to the sequence of selected shots, resulting in a trailer-like summary.

[2]Quite recently, a similar approach was adopted by an IBM research team, asked by the 20th Century Fox to develop a "cognitive movie trailer" for its upcoming film *Morgan*. For a preliminary, informal description of the method used, see http://asmarterplanet.com/blogs/think/2016/08/31/cognitive-movie-trailer/

2.1.2 Use-cases and evaluation

The sport meeting highlighting techniques described above implicitly target restricted use-cases, where users are assumed to be interested in specific, well-defined classes of events. Consequently, the detection of such salient events can be objectively evaluated in a fully automatic process, where a reference ground-truth is available. In (Chang et al., 2002) for example, four classes of baseball events are objectively regarded as highlights; similarly, in (Assfalg et al., 2003), pre-defined classes of soccer highlights, such as *shot on goal* or *turnover* are used as reference salient events; and the *play* events considered in the more generic framework introduced in (Li et Sezan, 2001) can still be defined in a relatively objective way. Highlight detection is then performed and evaluated as a categorization task, with standard evaluation metrics (precision, recall, F-score). The more generic framework introduced in (Hanjalic et Xu, 2005) for indirectly detecting sport highlights from external reactions is evaluated by qualitatively comparing reference soccer highlights to the subjective excitement, modeled from low-level features, they are likely to trigger.

Even when automatically selecting sequences for further insertion in action movie trailers, reference to ground-truth is still possible: (Smeaton et al., 2006) for instance evaluate the relevance of the automatically selected shots by reference to those inserted in the original, handmade movie trailers.

Nonetheless, as stated in (Truong et Venkatesh, 2007), for general-purpose highlighting techniques, "the ground-truth tends to be slightly subjective", and subjective evaluation based on user studies is preferable.

First, movie trailers, as actual commercial artifacts, respond to advertising purposes, and the associated use-case is well-defined: viewers have not already watched the video source, and the summary is expected to be enjoyable enough to make them want to view it. In this case, the resulting trailer is only a video highlight focusing on engaging scenes, and is in no way expected to fully cover the original movie plot. Accordingly, (Irie et al., 2010) design a comprehensive evaluation framework, fully consistent with the specific use-case scenario commercial trailers target: 20 participants were asked to rate trailers of 16 full-length movies. For each movie, 4 trailers, including baseline and reference ones, were generated, and the users were asked to rate them according to 3 criteria: *appropriateness* ("How closely did the trailer look to an actual trailer?"); *impact* ("How much were you impressed by this trailer?"); and *interest* ("How interested did you become to watch the original movie?"). What the authors call *impact* and *interest* specify the standard *enjoyability* criterion in a way that is fully consistent with the typical use of trailers.

However, according to some authors, highlighting summaries do not necessarily reduce to the same promotional purpose as movie trailers. The use-case described in (Ma et al., 2002) goes far beyond designing appealing video summaries: when deciding if a source video is worth watching, viewers are assumed to use the informativeness criterion and summaries should be informative enough to help them to make their decision. Accordingly, the authors evaluate their attention-based summarization framework by

asking 20 participants to rate summaries of various videos according to two criteria: *enjoyability* targets the engaging aspects of the summary, while *informativeness* aims at evaluating its ability to cover the whole source content. Both criteria are commonly used for subjective evaluation of automatically generated summaries (see for instance the evaluation of the saliency-based summarization framework introduced in (Evangelopoulos et al., 2013)).

2.1.3 Discussion

Markers of formal saliency are widely available in video streams, and can often be captured from relatively low-level features. Based on formal saliency, highlighting summaries can automatically be generated: such summaries focus on the most salient parts of the video source but they do not usually pretend to provide the users a comprehensive overview of the original content.

Consequently, we do believe that such summaries should not be evaluated with respect to their ability to fully cover the source content: the *informativeness* criterion sometimes used when evaluating video highlights is not fully consistent with the low-level features they commonly rely on, more likely to capture engaging than meaningful sequences. Instead, the evaluation of such summaries should be application-dependent. When the summary aims at capturing pre-defined salient events, for instance for replaying parts of sport meetings, the evaluation process can be fully automated; when such a notion is more subjective, as in movies, the evaluation should depend on the use-case considered. Movie trailers for instance correspond to well-defined use-case scenarios: they provide viewers with expectedly attractive previews that should in no way reveal the whole storyline.

Automatically building *synopses*[3], *i. e.* summaries that could ideally replace the original stream with minimum information loss, turns out to be much more challenging from a computational perspective: the content coverage objective remains dependent on effective content-modeling, in order to properly handle the so-called *semantic gap*.

2.2 Synopses based on content coverage

Semantic gap

The semantic gap, first exhibited in the context of content-based image retrieval (Gudivada et Raghavan, 1995), is sometimes mentioned as a major issue for automatic video summarization. In the video context, this concept usually denotes the mismatch between the computable and human representations of the video content: the computable low-level features turn out to be poorly connected to the way humans commonly understand the related content, and we are left with "bridging the gap" by constructing intermediary representations.

[3] Denoted *summary sequences* in (Tsoneva et al., 2007).

In particular, movies and TV series are denoted in (Tsoneva et al., 2007) as *narrative media*. According to the authors, unlike event-centralized media like TV news, narrative media do not have underlying structures that could be used during the summarization process. Instead, the summarization process must capture the dynamics of the story such media report, and rely in the first place in modeling plot continuity.

2.2.1 Features for content modeling

Two main types of features have been used so far for modeling the plot content and preserving the dynamics of movies and TV series: the first ones are based on textual content, while the others rely on the network of interacting characters.

Textual

(Smith et Kanade, 1997) is an early attempt to integrate language understanding techniques into video summarization frameworks. Highly compressed video skims of documentaries and TV news are generated from keyphrases detected within the audio transcript after applying within every scene the standard TF/IDF lexical weighting scheme.

(Tsoneva et al., 2007) extend *natural language processing* techniques to generate video summaries of narrative media. The authors focus on automatic generation of comprehensive summaries of TV series episodes that would ideally be content-covering enough to dispense viewers from watching the original source. The content coverage objective is achieved by preserving the narrative continuity of the story. To this purpose, the subscenes, as resulting from the joint use of the movie script and subtitles, are ranked, among other features, according to their ability to preserve the main storyline by carrying lexically meaningful information. The summarization framework is applied to three TV series episodes and result in 15-minute summaries with low compression rates.

Social Network Analysis

The benefits of *social network analysis* (SNA) for investigating the content of fictional works in general have recently been emphasized in several articles. Most focus on literary works: *dramas*, analyzed from the SNA perspective either in static (Moretti, 2011) or dynamic (Nalisnick et Baird, 2013) ways; novels, for instance in (Elson et al., 2010) (corpus of 19^{th}-century British novels), (Agarwal et al., 2012) (L. Caroll's *Alice in Wonderland*), (Rochat et Kaplan, 2014) (Rousseau's autobiographic text *Les Confessions*); comics, in (Alberich et al., 2002; Gleiser, 2007) (SNA-based analysis of Marvel Universe); or even mythology in (Mac Carron et Kenna, 2012) (epic texts)...

Interestingly, the shift from textual to social features for automatically investigating the plot of novels and dramas originated in the literary studies themselves. (Sparavigna, 2013) reports how Franco Moretti's *Stanford Literary Laboratory* contributed to introduce in literary studies quantitative approaches based on SNA. (Moretti, 2011) underlines and illustrates with Shakespeare's Hamlet the light SNA can shed on literary works, either plays or novels: by aggregating all character's interactions in a single net-

work, SNA helps to unveil some underlying patterns invisible to close textual reading, which remains dependent on the narrative sequentiality.

Inspired by the narrative charts manually drawn on the website XKCD[4], some authors make use of character interactions to introduce alternative ways of providing the users with concise representation of movie plots. As shown on Fig. 2.1 for the movie trilogy *The Lord of the Rings*, such narrative charts allow to visualize character interactions on a timeline, and provide, especially if interactive, meaningful representations of the plot as possible alternatives to extractive video summaries. An early attempt to automatically generate such narrative charts from annotated data can be found in (Ogawa et Ma, 2010). (Tapaswi et al., 2014) explore a way of drawing narrative charts of single TV series episodes from automatically estimated interactions based on scene detection and face identification. (Liu et al., 2013) design a comprehensive, interactive framework to generate and visualize narrative charts, but based on external, manually introduced information.

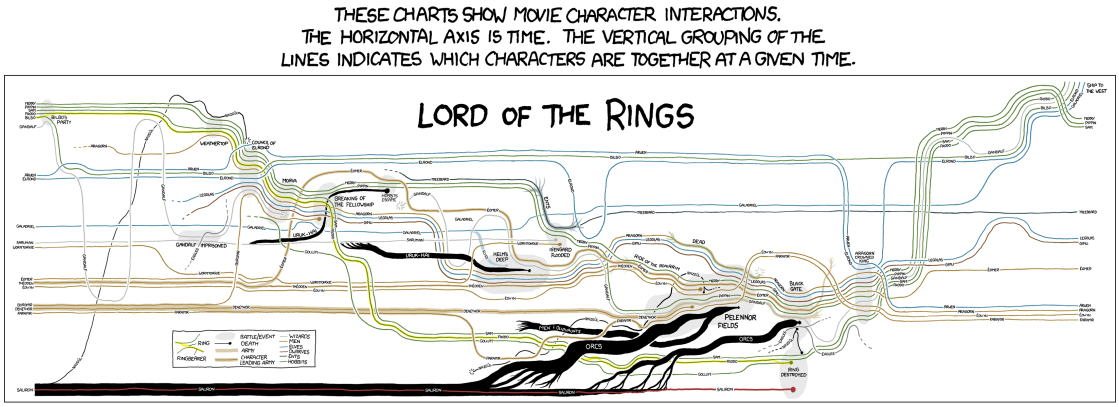


Figure 2.1 – *Excerpt from the* XKCD *comic Movie Narrative Charts (source:* https://xkcd.com/657/*).*

(Weng et al., 2007, 2009) explicitly aim at making use of SNA to bridge the gap between low-level video features and the way humans commonly understand movies: "when we watch a movie, what we really see are the *stories* derived from the action or interaction between characters". Accordingly, the authors make use of SNA to automatically analyze the plot of full-length movies and single episodes of classical TV series. They first build the social network of interacting characters upon scene co-occurrence. Based on the analysis of the resulting network into communities, the plot is then deinterlaced into major substories, and automatic detection of possible narrative breakpoints between scenes is performed.

(Tsoneva et al., 2007) make a first step towards using character-related information when designing their summarization framework. In addition to the textual features mentioned above, the ranking function they use to estimate the relevance of the candidate sequences targets characters evolving into stable communities, and important

[4] https://xkcd.com/657/

enough to be considered in the final summary. The protagonists' importance is estimated by the number of their narrative appearances, with a special interest for the first and last ones.

An additional step towards SNA-based video summarization approaches is performed in (Sang et Xu, 2010): inspired by the statement that "all films are about nothing – nothing but characters" (Monaco, 2000), the authors make use of SNA to design their summarization framework. Based on their possible social similarity, depending on the interacting characters, consecutive scenes are first clustered into substories as in (Weng et al., 2007). The relevance of the shots they contain is then estimated in a top-down fashion according to fully social criteria: the number of characters involved, the number of occurrences of leading characters and the amount of interaction between leading characters, with a special mention to opening/closing scenes and shots.

Further insight into SNA-based plot modeling is provided by the framework detailed in (Tsai et al., 2013) to automatically build semantic-preserving summaries of full-length movies. The approach relies on the extraction of a role-community network from the movie, where nodes represent groups of characters co-occurring in the same scenes, and arcs denote inclusion relationships between communities. The resulting communities are then clustered into narrative arcs, corresponding to narratively increasing groups of characters, and ranked according to their social impact. The generated summaries focus on non-redundant scenes covering the evolution of the most frequently activated narrative arc and show the progressive introduction in the story of the protagonists of the main subplot. Despite its limitations – plots with parallel narrative arcs as in TV serials cannot be properly handled – , the approach explicitly aims at generating summaries capable of preserving the narrative continuity and makes an additional step towards SNA-based video summarization frameworks.

2.2.2 Use-cases and evaluation

Consistently with the content-covering features they use in their respective summarization frameworks, (Tsoneva et al., 2007) and (Tsai et al., 2013) explore "replacement" use-cases: the increasing amount of available data and the time constraints of modern life are assumed to result in demands for substituting summaries.

The evaluation protocol designed by (Tsoneva et al., 2007) is quite consistent with such a use: the informativeness of the resulting summaries is objectively exhibited by evaluating human understanding of the movie content: 36 users unaware of the original source were asked to shortly describe the story of 9 TV series episodes after viewing 10/15-minute summaries, and to answer a few questions about the content. The answers were then rated independently by two judges, resulting in a comprehensive evaluation of the summary ability to provide the users with understandable content. Moreover, user satisfaction was also evaluated with respect to various criteria. For every episode, four summaries were generated, and distributed among the users such that none of them could see several summaries of the same episode.

Similarly, the evaluation protocol detailed in (Tsai et al., 2013) is consistent with the

"replacement" use-case: the generated summaries are expected to be semantically covering enough to replace the original source with minimum loss of information. The SNA-based framework they introduce to achieve such a content coverage purpose is evaluated by asking 36 participants to evaluate 15/20-minute summaries of 12 full-length movies. For each movie, four summaries are generated, including two baselines, that users are asked to rank according to 6 criteria. The first two criteria are the standard "informativeness" and "enjoyability", but most of the remaining questions aim at evaluating the semantic coverage of the generated summaries. The experiment is performed in uncontrolled conditions, where viewers were free to watch the summaries, along with the original movie, as many times as they wanted.

The use-case considered by (Sang et Xu, 2010) when evaluating their summarization approach is not as ambitious as the features they use: the content-related reference to the interacting characters is only expected to result in engaging summaries for promotional purposes. The experimental evaluation is then performed in quite a standard way: 5 users aware of the source content are asked to rate summaries of one full-length movie and three sitcom episodes, according to the informativeness and enjoyability criteria; the summaries are built for various skim ratios (ranging from 10% to 30%), and are compared to summaries based on a user-attention model not sensitive to the movie content.

2.2.3 Discussion

Automatic generation of substituting video summaries that could dispense to watch the original source turns out to be much more challenging than building trailer-like summaries based on formal saliency. For narrative media, such a content coverage objective requires the plot to be modeled and preserved in the final summary with minimum loss of information. Social network analysis, possibly in conjunction with natural language processing techniques, turns out to be well suited for such a plot-modeling purpose and replacement use-case scenario.

Symmetrically to the concluding remarks we made in Subsection 2.1.3, we do think that the *enjoyability* criterion for evaluation purposes does not suit content-covering synopses. Even though sequences inserted into the summary for semantic reasons may be accidentally appealing, the evaluation criteria must be kept consistent with the kind of features the summarizing algorithm relies on. *Informativeness* should then be defined as a relative evaluation metrics, and denote the ability of the summary to cover the source content with minimum loss of information. Informativeness may then be evaluated by the users themselves as in (Tsai et al., 2013), asked to carefully compare on their own the summary to the original source; or indirectly as in (Tsoneva et al., 2007), where the semantic self-sufficiency of the generated summaries is evaluated by rating the level of understanding of users unaware of the source content. Whatever the way informativeness is evaluated, careful attention must be paid to the use of baseline methods for comparison purposes: if the same user were to view several summaries of the same content, the viewing order may result in biased feedback.

2.3 Summary

Table 2.1 reports the main use-cases identified in this section for movie summarization, along with some of the mentioned works, where the evaluation protocol is carefully designed in accordance with the use-cases considered. The resulting summaries may correspond to actual use-cases, where handmade summaries are widely available (in bold), or to more hypothetical use-cases with fewer potential users (in italic). The use-cases mentioned depend on the user's level of familiarity with the original content, together with the type of features considered. Though Table 2.1 concentrates on movie summarization, it can be used for specifying any video summary, whatever the video source.

Table 2.1 *– Typology for movie summarization.*

		Features	
		Saliency-oriented	Content-oriented
Movie	Already watched	–	*Reminding summaries* (Tsoneva et al., 2007)
	Not watched yet	**Promotional trailers** (Irie et al., 2010) ...	*Substituting summaries* (Tsoneva et al., 2007) (Tsai et al., 2013)

As can be seen, promotional trailers based on saliency markers correspond to well-defined use-cases, both from the content producers' and viewers' perspectives. Content-based summaries of full-length movies are not so widespread, and respond to much more marginal needs: the reminding scenario briefly mentioned in (Tsoneva et al., 2007), where users would like to watch summaries to remember the source content, is much more likely to occur for TV serials with continuous plots than for full-length movies or even classical TV series with standalone episodes. The replacing scenario, where users would like to skim videos because of time constraints, both mentioned in (Tsoneva et al., 2007) and (Tsai et al., 2013), remains quite hypothetical for movies and is more likely to emerge for documentaries or TV news contexts where information is provided to potentially busy viewers: when we watch narrative films based on fictional screenplays, part of the experience consists in getting fully immersed in the plot.

In contrast, and as can be seen in Table 2.2, TV serials provide us with quite realistic use-cases.

Table 2.2 *– Typology for* TV *serial summarization.*

		Features	
		Saliency-oriented	Content-oriented
TV serial	Already watched	**Re-engaging summaries**	**Reminding summaries**
	Not watched yet	**Promotional trailers**	_

Saliency-oriented summaries of TV serials are expected to affectively re-immerse

the viewers into a plot they used to be familiar with, resulting in *re-engaging summaries*; *reminding summaries* based on content modeling are expected to help them to remember the plot content, and *promotional* trailers to make them want to watch the incoming episode/season. For all of these use-cases, handmade video summaries are widely available, in response to actual needs. The remaining case, where users would consult summaries to catch up with missed episodes, remains too marginal to be mentioned, as shown on Fig. 1.10 in Section 1.3.

As stated in Section 1.3, by building summaries that address radical drop-off in viewer's engagement between TV serials seasons, we explore the use of standard saliency-oriented features, but in quite a novel, realistic use-case scenario: instead of using such features to promote incoming episodes/season, we make use of these features to affectively revive the main storylines and make viewers want to know what comes next in new seasons of TV serials.

Furthermore, we explore plot modeling in quite a novel way: first, by considering the remembering use-case scenario, much more realistic for TV serials than for full-length movies; second, by exploring plot modeling at the unprecedented scale of dozens of consecutive episodes, where saliency-based techniques are likely to miss important narrative developments.

Finally, the validation of our summarization framework in real-world conditions relies on a large-scale user study, with many more participants than in the user studies supporting the works mentioned in this section[5].

[5]Such large-scale validation protocols in realistic conditions are sometimes denoted as *ecological*.

Chapter 3

Video segmentation Feature extraction

Contents

3.1 Introduction

The algorithms we use to automatically build video summaries of TV serials are dependent on various preliminary subtasks.

First, segmenting the video stream at different levels is required: the extractive video summaries we intend to generate consist of excerpts from the original stream, and the elementary units candidate for later insertion in the summaries need to be defined and automatically extracted; moreover, the extraction of some of the mid-level features we use when estimating the relevance of each candidate excerpt is dependent on the delimitation of some video sequences. We describe in Section 3.2 the methods and definitions we use for segmenting the video stream at different levels of granularity, from the atomic video frames natively available to larger, semantics-related, video sequences.

Secondly, automatic summarization rely on estimating the relevance of each candidate unit based on some features, either content or saliency-oriented. From Section 3.3 up to the end of this chapter, we detail the mid-level features, along with the feature extraction methods, we use for modeling both the style and content of TV serial narratives: the way we estimate *shot size* is described in Section 3.3. The method we use for tracking *background music* is described in Section 3.4. Both these style-related features are extracted by applying standard algorithms. Our own contributions to the *speaker diarization* task are detailed in Section 3.5.

3.2 From video frames to scenes

In this section, we define every video unit used in this work, along with the segmentation techniques we apply:

- Subsection 3.2.1 describes how video shots are retrieved from the sequence of video frames (atomic units), before Subsection 3.2.2 introduces the standard method we use for detecting similar, recurring shots.

- Based on recurring shots, two types of shots sequences are then considered: the first ones (Subsection 3.2.3) are based on fixed patterns of alternating recurring shots and are used when performing speaker diarization; the second ones, denoted *Logical Story Units*, either maximal or not, are introduced in Subsection 3.2.4, along with a novel extraction algorithm, as the basic candidate units for later insertion in the summaries.

- Scenes are the larger video units considered and are used in Chapter 4 when building the social network of interacting speakers.

3.2.1 Shot cut detection

In this subsection, we describe the standard algorithm we use for retrieving shot cuts from the most elementary visual units natively available in the video stream.

The whole video stream can be regarded as a sequence of fixed images (or video frames) displayed onscreen at a constant rate able to simulate for human eyes the continuity of motion. A video shot, as stated in (Koprinska et Carrato, 2001), is defined as an "unbroken sequence of frames taken from one camera". Shot boundaries can then be detected by comparing the current image to the next one: a substantial difference between two temporally adjacent images is indicative of a shot cut[1].

Global comparison of color histograms

We hypothesize a cut between two contiguous shots whenever the difference between two temporally adjacent images exceeds the *differentiation* threshold τ_1. For this comparison purpose, images are described by 3-dimensional histograms ($24 \times 8 \times 64$ bins) of the image pixel values in the HSV color space[2]. Comparison between images is then performed by evaluating the correlation between their respective color histograms. Fig 3.1 shows three consecutive video frames (top), along with their two-dimensional hue/saturation (bottom left) and one-dimensional value (bottom right) histograms.

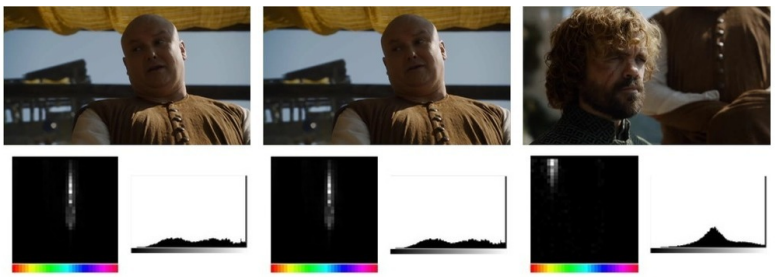


Figure 3.1 – *Sequence of three consecutive video frames (top), along with their two-dimensional hue/saturation (bottom left) and one-dimensional value (bottom right) histograms.*

As can be seen from Fig 3.1, whereas the histograms computed from the first two images, belonging to the same shot, remain well correlated, the histograms extracted from the last two ones, separated by a shot cut, look very different from each other and turn out to be poorly correlated, whatever the dimension of the HSV color space.

Block-based comparison of local color histograms

Nonetheless, different images might share the same global color histogram, resulting in false negatives when performing shot cut detection. The basic algorithm described and illustrated above is then further refined: information about the spatial distribution of

[1] Marginal in our corpus of TV serials, gradual transitions (*fades*) between shots are here discarded and we only take into account abrupt ones (*cuts*).

[2] The *Hue/Saturation/Value* color space, closer to the human perception of colors than the traditional RGB one, empirically turns out to be the most appropriate representation of colors in our corpus.

the colors over the image is reintroduced by splitting the image into 30 (6×5) blocks of equal size, each associated with its own HSV color histogram; block-based comparison between images is then performed as described in (Koprinska et Carrato, 2001), by averaging the correlations between the corresponding local histograms in both images.

Furthermore, in order to reduce the number of false detections that could possibly result from fast motion, we define a *similarity* threshold τ_2 to ensure that the sequence of frames surrounding two frames possibly separated by a shot cut remains quite stable. A shot boundary is then hypothesized between the frames f_2 and f_3 within the frame sequence $f_1 f_2 f_3 f_4$ if and only if the three following conditions stand:

$$\bar{r}(f_1, f_2) \geqslant \tau_2, \;\; \bar{r}(f_2, f_3) \leqslant \tau_1, \;\; \bar{r}(f_3, f_4) \geqslant \tau_2$$

where $\bar{r}(f_i, f_j)$ denotes the average correlation resulting from the block-based comparison of the local color histograms, as extracted from both video frames f_i and f_j. We empirically set both thresholds τ_1 and τ_2 to 0.7 and 0.8 respectively, after maximizing the evaluation metrics associated with the shot detection task on a development set.

Experimental results

The evaluation of the shot cut detection task relies on a standard F1-score (Boreczky et Rowe, 1996) based on recall (proportion of retrieved cuts among the reference ones) and precision (proportion of relevant cuts among the retrieved ones). The results on a development set (denoted DEV) consisting of 6 episodes, along with the results obtained on a test set of 3 episodes, are reported in Table 3.1. In both sets, the three TV serials of our corpus are equally represented.

***Table 3.1** – Results obtained for shot cut detection*

TV serial	episode	precision	recall	F1-score
BB	S01E01	0.98	0.88	0.93
	S01E02	1.00	0.98	0.99
GoT	S01E01	0.98	0.96	0.97
	S01E02	0.98	0.98	0.98
HoC	S01E01	0.99	0.98	0.99
	S01E02	0.98	0.98	0.98
avg. DEV		0.99	0.96	**0.97**
BB	S01E03	1.00	0.97	0.98
GoT	S01E03	0.99	0.98	0.98
HoC	S01E03	0.99	0.99	0.99
avg. TEST		0.99	0.98	**0.98**

As can be seen in Table 3.1, the scores obtained are quite high, and even improve on the test set. Shot cut detection is indeed considered as a well-solved task in the image processing field, and the basic thresholding techniques we described above are quite sufficient for our purpose. Nonetheless, some cuts were missed in BB, S01E01,

resulting in a lower recall than in the other episodes: most of these missed cuts belong to the same single scene, where unusual editing rules were applied by the filmmaker.

3.2.2 Detection of recurring shots

In this subsection, we show how the same visual features can be used for detecting similar, recurring shots in the video stream.

The same representation of images from local color histograms and the same basic thresholding techniques as the ones used in Subsection 3.2.1 make it straightforward to cluster recurring, similar shots, *i. e.* shots capturing the same scene settings, as being filmed from the same point of view. Figure 3.2 shows an example of a shot sequence involving two recurring shots, respectively labeled l_1 and l_2, within a dialogue.

Figure 3.2 – *Example of shot sequence ...$l_1 l_2 l_1 l_2$... with two recurring shots, respectively labeled l_1 and l_2.*

For such a clustering purpose, we use a *similarity* threshold τ_3, defined as the minimum average correlation required between the local color histograms of the previous shots' last frame and the current shot's first frame. Each shot is tested as possibly paired with past ones as the same recurring shot within an observation window of 30 consecutive shots. We experimentally set the value of τ_3 to 0.6 by maximizing on the same development set as in Subsection 3.2.1 the evaluation metrics when detecting similar, recurring shots.

Experimental results

We evaluate the resulting shot clusters by applying an adapted F1-score: for each shot, the list of shots hypothesized as similar to the current one is compared to the reference list of similar shots; if both lists intersect in a non-empty set, the shot is considered as correctly paired with its list. The results obtained on the DEV and TEST sets are reported in Table 3.2.

Once again, the results, despite the basic thresholding technique we use, remain remarkably stable when obtained on the test set. When setting the threshold τ_3, we generally preferred precision to recall, resulting in shorter, but more reliable, shot sequences of the type we describe in the two following subsections.

Table 3.2 *– Results obtained for shot similarity detection*

TV serial	episode	precision	recall	F1-score
BB	S01E01	0.87	0.81	0.84
	S01E02	0.89	0.84	0.86
GoT	S01E01	0.85	0.84	0.84
	S01E02	0.88	0.89	0.89
HoC	S01E01	0.91	0.93	0.92
	S01E02	0.96	0.90	0.93
avg. DEV		0.89	0.87	**0.88**
BB	S01E03	0.82	0.84	0.83
GoT	S01E03	0.91	0.80	0.85
HoC	S01E03	0.98	0.96	0.97
avg. TEST		0.90	0.87	**0.88**

3.2.3 Dialogue shot patterns

Once shots are extracted and similar ones are detected, we can automatically retrieve a first kind of video sequence typical of short dialogues, which turns out to be especially useful when performing speaker diarization.

Filming dialogues

From the filmmaking perspective, dialogues require the *180-degree rule*[3] to be respected in order to keep the exchange between characters natural enough: so that both speakers seem to look at each other when successively appearing onscreen, the first one must look right and the second one must look left. To achieve this, two cameras must be placed along the same side of an imaginary line connecting them. Fig. 3.3 illustrates the application of the 180-degree rule when filming an interaction between two characters.

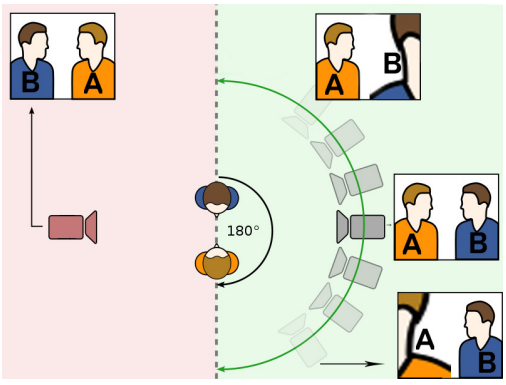


Figure 3.3 *– Illustration of the 180-degree rule.*

[3]See for instance http://learnaboutfilm.com/film-language/sequence/180-degree-rule, or in French https://www.youtube.com/watch?v=-f_8jY2ilpY.

The application of the 180-degree rule results in a specific visual pattern made of two alternating, recurring shots, highly typical of dialogues. Fig. 3.2 illustrates such a pattern. We now formalize the way such dialogue patterns can be captured.

Basic dialogue pattern

Let $\Sigma = \{l_1, ..., l_m\}$ be a set of possible shot labels, two shots sharing the same label if they are similar to one another. The following regular expression $r(l_1, l_2)$ corresponds to a subset of all the possible shot sequences $\Sigma^* = \bigcup_{n \geqslant 0} \Sigma^n$:

$$r(l_1, l_2) = \Sigma^* l_1 (l_2 l_1)^+ \Sigma^* \tag{3.1}$$

The set $\mathcal{L}\left(r(l_1, l_2)\right)$ of sequences captured by the regular expression 3.1 corresponds to shot label sequences containing one occurrence of l_2 inserted between two occurrences of l_1, with a possible repetition of the alternation $(l_2 l_1)$, whatever the previous and following shot labels. Such a regular expression formalizes the "two-alternating-and-recurring-shots" basic pattern described above as typical of dialogue sequences involving two characters. For example, the sequence of shot labels illustrated on Fig. 3.2 match the regular expression 3.1.

Moreover, a movie can be described as a finite sequence $\mathbf{s} = s_1 ... s_k$ of k shot labels, with $s_i \in \Sigma$. The set of patterns $\mathcal{P}(\mathbf{s}) \subseteq \Sigma^2$ associated with the movie shot sequence $\mathbf{s}$ can be defined as follows:

$$\mathcal{P}(\mathbf{s}) = \left\{(l_1, l_2) \in \Sigma^2 \mid \mathbf{s} \in \mathcal{L}\left(r(l_1, l_2)\right)\right\} \tag{3.2}$$

In other words, the set of patterns $\mathcal{P}(\mathbf{s})$ contains all the pairs of shots alternating with each other according to Rule 3.1.

Extended dialogue pattern

In order to increase the coverage of the patterns included in $\mathcal{P}(\mathbf{s})$ and reduce their sparsity, two extensions of the condition 3.1 are introduced:

1. In addition to Rule 3.1, isolated expressions of the two alternating shots of the form $(l_1 l_2 | l_2 l_1)^+$ are taken into account, increasing the total amount of speech captured by the patterns.

Figure 3.4 – *Shot sequence ...$l_1 l_2 l_1 l_3 l_1$... at the boundary of two adjacent patterns (l_1, l_2) and (l_1, l_3) with one shot in common.*

2. Moreover, the number of patterns is reduced while the average pattern coverage is increased by iteratively merging in a new pattern two patterns (l_1, l_2) and (l_1, l_3) with

at least one label in common. As showed on Fig. 3.4, such situations frequently occur within dialogues when one of the speakers (here the one appearing on the shots l_2 and l_3) is being alternatively filmed from two distinct cameras. The resulting pattern gather all the utterances $\mathbf{u}(l_1, l_2)$ and $\mathbf{u}(l_1, l_3)$ covered by the merged patterns.

Dialogue patterns: coverage

Table 3.3 reports the total coverage of the patterns extracted from the 9 TV serial episodes mentioned in Subsections 3.2.1 and 3.2.2, expressed in % as the ratio between the amount of speech covered by the patterns and the total amount of speech. The average duration of the speech covered by each pattern is also indicated, as well as the average number of speakers by pattern. These data are both computed by applying the basic version of the regular expression r, as given in Equation 3.1, and by using its extended version.

Table 3.3 – *Dialogue patterns and speech: statistical data*

	coverage (%)	**spch/patt (s.)**	**# of spk/patt**
r	49.51	11.07	1.77
ext. r	**51.99**	**20.90**	**1.86**

As indicated in Table 3.3, the extracted visual patterns cover in average a bit more than half (51.99%) of the total amount of speech contained in the 9 considered TV serial episodes. As expected, the average number of speakers by pattern remains quite low: 1.86, with the following distribution (not reported in Table 3.3): 69.85% of the patterns contain 2 speakers, 8.09% three and 22.06% only one. However, most of these one-speaker patterns correspond to short scenes, where the chance that one of the speakers remains silent increases.

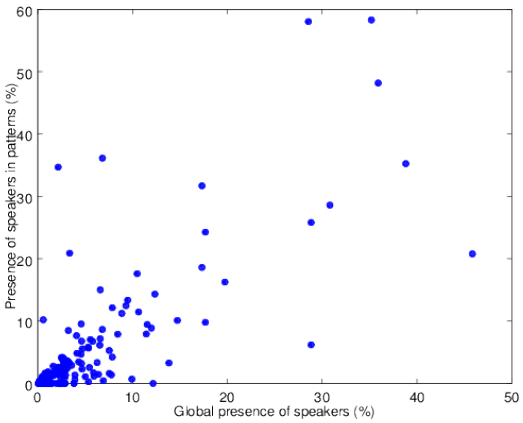


Figure 3.5 – *Speaker involvement within dialogue patterns, plotted as a function of speaker global involvement.*

Figure 3.5 shows the ability of such visual patterns to capture the main characters in the 9 considered episodes, by plotting the speakers' verbal involvement in patterns (expressed in %) as a function of their global involvement (in %). As can be seen, ac-

tive speakers are well represented in the considered patterns: 97.96% of the characters speaking at least 5% of the time are involved in such patterns.

In summary, the visual patterns described in this subsection provide us with short dialogue sequences with a small number of speakers involved, and are used in Section 3.5 as a reliable basis when locally performing speaker diarization, either in a mono-modal (Subsection 3.5.3), or multi-modal (Subsection 3.5.4) way.

3.2.4 Logical Story Units

In this subsection, we make an additional step towards larger video sequences and we define the video sequences we regard as the basic candidate units for potential insertion in the summary, along with a novel algorithm for extracting them.

Logical Story Units: definition

Though built upon sophisticated editing rules, the "official " video recaps of TV serials rarely concatenate isolated shots extracted from different parts of the original stream. Instead, the basic unit used in such summaries is typically a short sequence of about 10 seconds consisting of a few consecutive shots. Such sequences are usually selected not only because of their semantic relevance, but also because of their semantic cohesion and self-sufficiency. From a computational perspective, identifying such sequences in the video stream as potential candidates for later insertion in the final summary remains tricky.

Nonetheless, the stylistic patterns widespread among filmmakers, such as the application of the 180-degree rule illustrated on Fig. 3.3, are particularly relevant, because they are often used to emphasize the semantic consistency of these sequences. Because multiple cameras may be used to film the same character within dialogues, the application of the 180-degree rule often results in more complex patterns than the basic case of two alternating, recurring shots described in Subsection 3.2.3: several sets of recurring shots may overlap each over with single, non-recurring shots in-between, resulting in longer patterns well suited for segmenting movies into consistent narrative episodes. In (Hanjalic et al., 1999), the authors introduce the notion of *Logical Story Units* (LSUs) to denote such intertwined sequences of recurring shots, along with a method for automatically extracting them. LSUs are well suited for capturing consistent scene settings, especially within dialogues: as long as the video stream contains recurring shots, the captured events occur both at the same place and over a continuous period of time.

Indeed, both Fig. 3.2 and Fig. 3.4, respectively captured by the basic regular expression 3.1 and its extended version, illustrate special cases of LSUs. The notion of LSU is actually much more general and covering than the limited shot sequences described in Subsection 3.2.3: the only presence of the surrounding shot l_1 in the shot sequence shown on Fig. 3.4 is sufficient to define this sequence as an LSU, whatever the three intermediate shots, whereas it would not be captured by Rule 3.1 if no recurring shots occurred in-between. We therefore choose LSUs as the general candidate units used when building the summaries of TV serials.

Extraction algorithm

The standard graph-based algorithm introduced in (Yeung et al., 1998) for segmenting the shot sequence into LSUs is fast, but may be tricky to implement: each cluster of similar shots is represented as a graph node, and directed links are introduced between two nodes if at least one shot belonging to the first corresponding cluster temporally follows one of the shots that belong to the second one. The LSUs boundaries then correspond to bridges (sometimes denoted cut edges) in the resulting so-called *Scene Transition Graph*, and linear-time algorithms are known for detecting such bridges in graphs.

We introduce here a novel, alternative matrix-based algorithm to automatically detect LSU boundaries: though computationally more expensive, such an algorithm has the advantage of relying on more simple data structures that make it easy to implement. The resulting shot sequences are strictly the same as when applying the standard algorithm: both only depend on the reliability on the shot similarity detection step.

Once performed, shot similarity detection results in a symmetric similarity matrix $\mathcal{S}$, where $s_{i,j}$ is set to 1 if the i^{th} and j^{th} shots are considered as similar, and to 0 otherwise. Provided that the shots are chronologically ordered, such a representation constitutes a straightforward way of automatically detecting the LSU boundaries. For instance, for the five shots included in the sequence shown on Fig. 3.4, the similarity matrix $\mathcal{S}$ is filled as follows:

$$\mathcal{S} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & \boxed{1} & 0 & 1 \\ 0 & 0 & 0 & \boxed{1} & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \qquad S^{(4)} \tag{3.3}$$

The k^{th} shot ($1 < k < n$, where n is the total number of shots within the sequence) is strictly included in one LSU if surrounded by at least two occurrences of the same recurring shot: in the matrix $\mathcal{S}$, such a statement is equivalent to the fact that the double sum $S^{(k)} := \sum_{(i>k, j<k)} s_{i,j}$ is greater or equal to 1. For instance, in Equation 3.3, the terms of the sum $S^{(4)}$ are included in the dashed red box. The fact that $S^{(4)} \geqslant 1$ means that the 4^{th} shot (solid red box) is surrounded by at least two occurrences of the same recurring shot, the first one occurring after ($i > 4$) and the second one before ($j < 4$) the 4^{th} position, and strictly belongs to one LSU.

The LSU boundaries can then be easily deduced from the two quantities $S^{(k)}$ and $S^{(k-1)}$, ($1 < k < n$, with $S^{(1)} := 0$) according to the two following rules:

- If $S^{(k-1)} = 0$ and $S^{(k)} \geqslant 1$, the $(k-1)^{\text{th}}$ shot is the beginning of a new LSU.
- Conversely, if $S^{(k-1)} \geqslant 1$ and $S^{(k)} = 0$, the k^{th} shot is the end of the previous LSU.

Furthermore, the double sum $S^{(k)} := \sum_{(i>k, j<k)} s_{i,j}$ does not need to be computed for each $k = 2, ..., (n-1)$ but can be recursively deduced from the previous quantity $S^{(k-1)}$ according to the following relation:

$$S^{(k)} = S^{(k-1)} - \sum_{j<k-1} s_{k,j} + \sum_{i>k} s_{i,k-1} \tag{3.4}$$

For example, in Equation 3.3, the quantity $S^{(4)}$ (sum of the coefficients inside the dashed red box) can be recursively obtained from the the quantity $S^{(3)}$ (sum of the coefficients inside the dashed blue box) as follows:

$$S^{(4)} = S^{(3)} - (s_{4,1} + s_{4,2}) + (s_{5,3})$$

By construction, the value of the coefficient $s_{k,k-1}$ is equal to 0 (two consecutive shots cannot be similar to each other) and is ignored when recursively updating the quantity $S^{(k)}$ from $S^{(k-1)}$ according to Equation 3.4.

The algorithm we use requires two nested loops over every shot, resulting in a time complexity in $O(n^2)$. The full method is summarized in Algorithm 1.

Algorithm 1 LSUs EXTRACTION

Require: $(s_{i,j})_{i,j=1,...,n} \in \{0,1\}$
$L \leftarrow \emptyset$
$S^{(1)} \leftarrow 0$
for $k \leftarrow 2$ to $(n-1)$ **do**
  $S^{(k)} \leftarrow S^{(k-1)}$
  for $j \leftarrow 1$ to $(k-2)$ **do**
    $S^{(k)} \leftarrow S^{(k)} - s_{k,j}$
  end for
  for $i \leftarrow (k+1)$ to n **do**
    $S^{(k)} \leftarrow S^{(k)} + s_{i,k-1}$
  end for
  if $S^{(k-1)} = 0$ **and** $S^{(k)} \geqslant 1$ **then**
    $start \leftarrow (k-1)$
  end if
  if $S^{(k-1)} \geqslant 1$ **and** $S^{(k)} = 0$ **then**
    $L \leftarrow L \cup (start, k)$
  end if
end for
if $S^{(n-1)} \geqslant 1$ **then**
  $L \leftarrow L \cup (start, n)$
end if
return L

LSUs: coverage

Table 3.4 reports statistical data about LSUs, as extracted from the 9 TV serial episodes mentioned in Subsections 3.2.1 and 3.2.2: their coverage is expressed in % of the total number of shots; the average number of shots by LSU as well as the maximum number of shots contained in one LSU are respectively reported in the 5th and 6th columns.

Table 3.4 – LSUs*: statistical data*

TV serial	episode	#	cov. in %	avg. # of shots	max. # of shots
BB	S01E01	50	76.89	12.58	77
	S01E02	43	70.27	9.07	46
	S01E03	33	75.60	14.18	77
GoT	S01E01	56	74.03	12.98	91
	S01E02	43	78.26	15.06	83
	S01E03	44	81.47	16.38	57
HoC	S01E01	56	63.28	8.46	45
	S01E02	37	59.18	10.97	42
	S01E03	40	79.68	16.28	61

As can be seen in Table 3.4, LSUs turn out to be much more covering and longer than the basic patterns described in Subsection 3.2.3: every LSU contains in average 13 shots, for a total coverage of about 75 %.

Nonetheless, many of the LSUs remain far too long to be directly inserted into TV serial summaries: the longest one in Table 3.4 consists of more than 90 shots (GoT, S01E01), and would be far too long to be directly included into a summary of reasonable length.

Recursive extraction of non-maximal LSUs

In order to get shorter candidate sequences without loosing the semantic consistency of LSUs, we recursively apply Algorithm 1 within each LSU to obtain more elementary, not maximal, LSUs.

During the extraction process, we put both lower and upper bounds on the duration of the candidate LSUs (either maximal or more elementary). We constrain every final candidate LSU to contain at least 3 shots, for a duration ranging from 5 to 15 seconds, close to the duration of the sequences inserted into handmade video summaries of TV series.

3.2.5 Scenes

Definition and coverage

Scenes are the largest video units we consider in this work: similarly to the rule of the three unities classically prescribed for dramas, a scene in a movie is defined as a

homogeneous sequence of actions occurring at the same place, within a continuous period of time. The whole video stream can be partitioned into scenes, which usually turn out to be larger shot sequences than LSUs.

Scene boundaries for interaction estimate

We make use of scenes as the most covering units when estimating in Chapter 4 the interacting speakers from the sequence of speech turns. When using this sequential estimate of speaker interactions, scene boundaries are needed: as shown on Fig. 3.6, a typical sequence of three different speakers is usually found at scene boundaries, the first one for instance belonging to the first scene, and the two other ones to the second scene. The only sequence of speech turns could then lead to introduce irrelevant interactions (here between the first and second labeled speakers) if the scene boundaries were not introduced to split the whole sequence of speech turns into distinct subsequences.

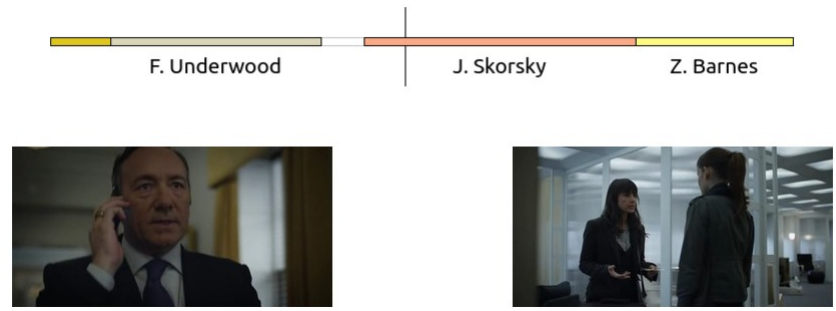


Figure 3.6 – *Sequence of speech segments distributed over a scene boundary (vertical line on top).*

LSUs for scene boundaries detection: evaluation

As stated above, scenes remain too large units to be automatically detected by extracting LSUs, and the scene boundaries detection task generally requires multi-modal approaches, possibly in addition to image-based LSU extraction as in (Ercolessi et al., 2011) and (Bredin, 2012). Table 3.5 and Table 3.6 evaluate the ability of image-based LSUs to match reference scene boundaries in 9 episodes from our corpus; in order to fully partition from LSU boundaries the whole video stream into scenes, the sets of shots occurring between two consecutive LSUs that are not adjacent are hypothesized as (transition) scenes.

The evaluation metrics used in Table 3.5, based on a standard F1-score, requires the hypothesized scene boundaries to match exactly the reference ones, whereas Table 3.6 relies on the less penalizing evaluation metrics described in (Vendrig et Worring, 2002): coverage (ideally maximal) evaluates the ability to cover the whole scenes and penalizes over-segmenting methods; overflow (ideally minimal) captures the ability of the method to detect scene changes and penalizes under-segmenting methods. Once complemented to 1, overflow is combined with coverage in a standard F1-score.

Table 3.5 *– Image-based* LSUs *for scene segmentation: evaluation (1)*

TV serial	episode	precision	recall	F1-score
BB	S01E01	0.24	0.76	**0.36**
	S01E02	0.14	0.56	**0.22**
	S01E03	0.01	0.55	**0.17**
GoT	S01E01	0.20	0.47	**0.28**
	S01E02	0.23	0.64	**0.34**
	S01E03	0.27	0.72	**0.39**
HoC	S01E01	0.23	0.50	**0.32**
	S01E02	0.30	0.46	**0.37**
	S01E03	0.50	0.73	**0.59**
average		0.18	0.60	**0.34**

Table 3.6 *– Image-based* LSUs *for scene segmentation: evaluation (2)*

TV serial	episode	coverage	overflow	F1-score
BB	S01E01	0.65	0.07	**0.76**
	S01E02	0.56	0.12	**0.69**
	S01E03	0.47	0.05	**0.63**
GoT	S01E01	0.76	0.23	**0.77**
	S01E02	0.74	0.02	**0.83**
	S01E03	0.77	0.03	**0.86**
HoC	S01E01	0.68	0.12	**0.76**
	S01E02	0.83	0.24	**0.79**
	S01E03	0.92	0.13	**0.89**
average		0.71	0.11	**0.78**

As can be seen in Table 3.5, recall is on average much higher than precision (0.60 *vs.* 0.18). Similarly, Table 3.6 reports a very low overflow (0.11) while coverage remains at a relatively low level (0.76): whatever the evaluation metrics, LSU-based segmentation into scenes tends to introduce many irrelevant scene boundaries while missing very few of them, resulting in over-segmenting the video stream. Indeed, LSUs are made of sequences of recurring shots, which ensures to capture the scene spatial and temporal continuity, resulting in high recall (or, equivalently, low overflow). Nonetheless, the whole sequence of actions associated to a specific scene remains usually much larger and is not likely to be surrounded by the same opening/concluding shot, which results in low precision (or, equivalently, low coverage), with many irrelevant boundaries. LSU detection can therefore not be used on its own as a reliable scene detection technique. As a consequence, we decided to manually introduce scene boundaries in every TV serial episode of our corpus. The needed annotation time surprisingly turned out to be quite short (about $1/10^{\text{th}}$ of real time).

3.2.6 Summary

In this section, we defined every video unit used in this work, from the smallest to the largest ones, along with the segmentation techniques on which they are based: video shots, possibly similar (Subsections 3.2.1 and 3.2.2); sequences of recurring shots based on fixed patterns (Subsection 3.2.3), used for later performing speaker diarization; Logical Story Units (Subsection 3.2.4), either maximal or not, defined as the basic candidate units for insertion into extractive summaries; scenes (Subsection 3.2.5), manually introduced for estimating interactions between speakers.

We now turn to the features we use for performing the summarization task, either saliency-oriented (changes in shot size and possible background music), or content-related (speakers involved in the audio stream).

3.3 Shot size

The semantics of most movie sequences does not only depend on their objective contents, but also on the way they are filmed and edited. For instance, the importance of a specific sequence in the plot, though primarily dependent on the content of the associated event, is usually emphasized by some stylistic patterns widely used in filmmaking. We first focus on a commonly used stylistic pattern, expected to provide us with a straightforward way of distinguishing salient sequences: the size of the shots.

From the film grammar perspective, shots sizes are distributed over a finite set of about 9 discrete classes, from extreme long shots to extreme closeups[4]. Fig. 3.7 shows examples of 3 shot size classes, extracted from a single video sequence with increasing shot size. However, rather than using such a set of discrete classes, we estimate shot size on a continuous scale ranging from 0 to 1, as the ratio between the apparent size of the face(s) detected onscreen and the image size.

Figure 3.7 – *3 shot size classes: from left to right,* mid shot, medium closeup, closeup.

In order to perform face detection, we use the face detector described in (Felzenszwalb et al., 2008; Girshick et al., 2012): we run the face detector on a sample of 5 video frames for each shot. The 5 frames are uniformly distributed over time within the shot. Such a sample size of 5 frames aims both at keeping the computation time reasonable when performing face detection, and at facing the issue of the characters

[4] http://learnaboutfilm.com/film-language/picture/shotsize

adopting various poses during a single shot, which is likely to cause false negatives. For each of the 5 frames, we retain, if any, the largest face box.

We then compute the absolute shot size as the median height of all the face boxes detected over the 5 frames and we finally express it as a proportion of the video frame height. Using the median rather than the mean value prevents shot size from being biased by very large or very small false positives, usually hypothesized for a single frame only.

Fig. 3.8 shows a sequence of four consecutive shots along with the face boundaries as automatically detected; on the top of the figure, the height of the gray rectangles corresponds to the resulting shot size, as a proportion of the frame height.

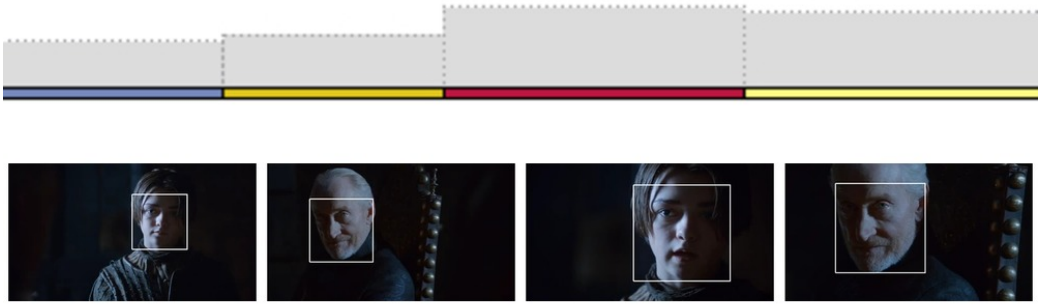


Figure 3.8 – *Shot sequence with face boundaries as automatically detected. On the top part, the height of the gray rectangles corresponds to the shot size, as a proportion of the video frame height.*

For both of the characters shown on the figure, shot size is increasing, from medium closeups to closeups, with the obvious intention to make the viewer focus on the last two shots, likely to be semantically salient. In Chapter 5, we make use of shot size as a critical stylistic feature when estimating the relevance of each candidate LSU for possible insertion in the summaries.

3.4 Background music

The second stylistic feature we consider is music, commonly used by filmmakers to support and emphasize moving sequences and to control, as detailed in (Roth, 2013), the associated mood[5]. Nonetheless, because we focus on dialogue sequences to build our summaries (see Chapter 5), we are specifically interested in detecting *background* music, likely to overlap with speech. We therefore perform such a music tracking task as a prediction task over a continuous variable corresponding to the average music ratio over a time-window (average musicality), rather than as a classification task.

[5] http://learnaboutfilm.com/film-language/sound and, in French https://www.youtube.com/watch?v=erQVzr1Xbko&list=PL0416194348A330A5&index=2

Though stated as a "hard" task in (Giannakopoulos et al., 2008), music tracking in movies can be performed by using the so-called 12-dimensional "chroma vectors" (or chromas), which contain the distribution of the audio signal over the twelve notes of the octave. Fig. 3.9 shows three representations of the audio signal, as extracted from a 57-second excerpt of *House of Cards*: the audio waveform (top), the corresponding spectrogram (evolution of the spectrum over time, bottom left) and chromagram (evolution of the chroma component values over time, bottom right).

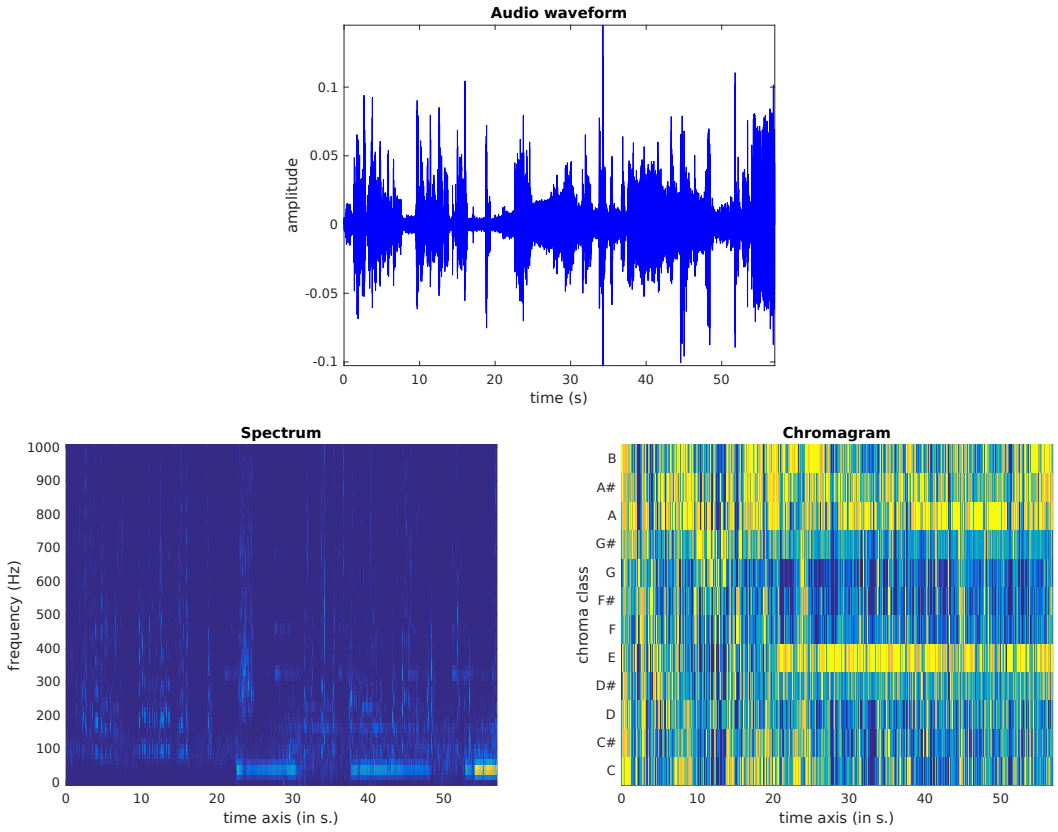


Figure 3.9 – *Audio waveform (top) from a 57-second excerpt of* House of Cards, *with background music (3 piano chords) overlapping with speech after the 21st sec.; corresponding spectrogram (bottom left) and chromagram (bottom right).*

Even when overlapping with speech, the starting point of the background music (just after the 21st sec.) can easily be detected from a visual inspection of both the spectrogram and chromagram: on the one hand, the frequencies look less uniformly distributed for music than for speech (high variance in the values of the chroma components considered statically); on the other hand, the frequencies are much more stable over time for music than for speech (low variance of every chroma component over time).

In order to capture these two features for the chroma vectors, we implemented with the MATLAB MIRtoolbox (Lartillot et al., 2008) the algorithm detailed in (Giannakopou-

los et al., 2008), resulting in two min-max normalized measurements of musicality over time: on the left-hand part of Fig. 3.10, both normalized static (expected to be high for music) and dynamic (expected to be low for music) spread of the chroma values are plotted as functions of time in the excerpt of *House of Cards* described above; the right-hand part of the figure shows their difference, as a single musicality measure.

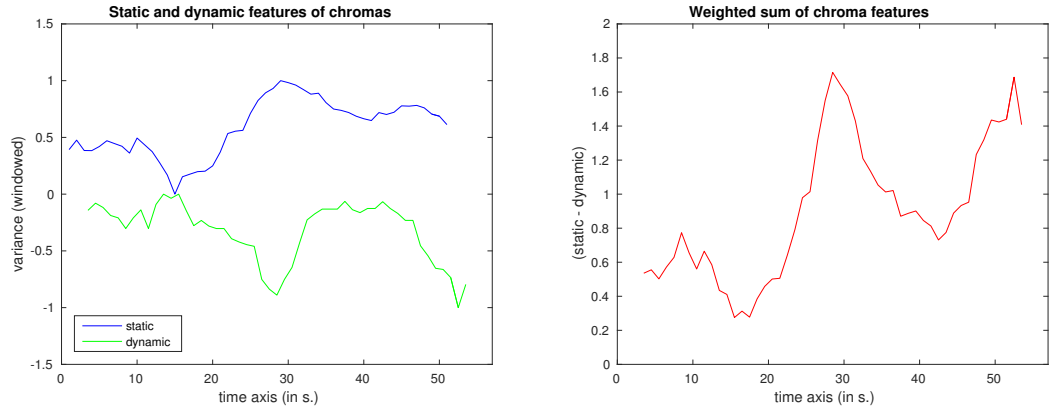


Figure 3.10 – *Static and dynamic spread of the chroma vectors components plotted as functions of time (left) in a 57-second excerpt of* House of Cards, *along with their difference (right).*

As can be seen from Fig. 3.10, the estimated musicality looks much higher in the second part of the excerpt (from the music starting point after the 21st sec.), where speech overlaps background "music" (actually a few piano chords) than in the purely-spoken first part. In Chapter 5 we make use of this chroma-based measure of musicality as a second clue, in addition to shot size, to discriminate stylistically salient sequences.

3.5 Speaker diarization

3.5.1 Introduction

Speaker diarization (denoted hereafter SD) consists in assigning the spoken segments of an audio stream to their respective speakers, without any prior knowledge about the speakers involved nor their number. Most state-of-the-art systems rely on a two-step approach, performing first speech turn detection followed by single-speaker segment clustering. This last stage is usually based on hierarchical clustering (Evans et al., 2012) and, more recently, mathematical programming (Dupuy et al., 2012, 2014), (Bredin et Poignant, 2013).

SD systems were first applied to audio-only streams produced in adverse but controlled conditions, such as telephone conversations, broadcast news, meetings... More recently, SD was extended to video streams, facing the critical issue of processing contents produced in uncontrolled and variable environments.

In (Clément et al., 2011), the authors apply standard SD tools to the audio source of various kinds of video documents. The reported results exhibit Diarization Error Rates (DER) much higher than for the classical application fields. The most dramatic decrease in performance is observed when the SD systems are applied to cartoons and movie trailers: among the possible reasons involved, the authors notice the high number of speakers involved in these streams, as well as the high variability of the acoustic environment (speech and music segments overlapping each other, sound effects). Moreover, as in most of the previous works on audiovisual SD, the diarization problem is here addressed by applying audio-only systems to the audio channel of videos, without any integration of the video-related features that could help the diarization system. Similar high error rates ($\simeq$ 70%) are reported in (Ercolessi, 2013) when applying a standard audio-only speaker diarization tool to the audio output of standalone episodes of TV series.

However, some recent works focus on multi-modal approaches for performing speaker segmentation of video streams: in (Friedland et al., 2009b), the authors evaluate a method based on early fusion of audio and video GMMs, and a classical BIC-based agglomerative process on the resulting two-channel information stream. This technique is evaluated on the AMI corpus (Carletta et al., 2005), which consists of audiovisual recordings of four participants playing roles in a meeting scenario. In (Bendris et al., 2013), the authors make use of both face clustering and speaker diarization to perform face identification in TV debates: face clustering and speaker diarization are first processed independently. Then, the current speaker is identified by selecting the best modality. Finally, local information about the current speaker identity is propagated to the whole cluster of the corresponding utterances. In (Bredin et Poignant, 2013), the authors make use of an intermediate fusion approach to guide speaker diarization in TV broadcast by adding to the set of speech turns new instances: the names written on the screen while a guest or a reporter is introduced as well as the corresponding identities. Adding such instances constrain the clustering process and result in purer classes of speakers.

Finally, audio-based SD has already been applied in (Bredin, 2012) to TV series, but as a mean among other modalities to structure its contents.

In this section, we introduce a novel SD framework well suited to TV serials, and more generally to fictional films. Such a framework is expected to provide, in addition to the stylistic features described in Sections 3.3 and 3.4, content-oriented information about the speakers involved.

Applying a SD system to TV serials, where the speaker number is generally higher than in full-length movies, is expected to be quite challenging. Nevertheless, as detailed in Section 3.2, TV serials, like any fictional film, are built upon filming techniques, such as the "180-degree" rule, that result in formal regularities at a visual level. From the dialogue shot patterns described in Subsection 3.2.3, we then propose to split the speaker diarization process into two steps when applied to TV serials: the first one consists in a step of local speaker diarization performed within sequences visually detected as dialogues; the next one consists in a step of constrained clustering that aims at retrieving the recurring speakers from one scene to the other: when performing this second

clustering step, we prevent the speakers locally hypothesized as different from being merged into the same global cluster (*cannot-link* clustering constraint).

Such a two-step clustering process is somehow related to what is denoted in (Tran et al., 2011) as the "hybrid architecture" in the cross-show speaker diarization context. In cross-show SD, diarization is achieved on a set of shows originating from a same source and containing possibly recurring speakers. The shows are first processed independently, before the resulting hypothesized speakers are clustered in a second stage.

We detail two possible methods to locally perform the first speaker diarization step:

1. The first one, described in Subsection 3.5.3, is based on the publication (Bost et Linares, 2014) and classically relies on audio-only features.
2. The second, alternative local speaker diarization framework, described in Subsection 3.5.4, is based on the publication (Bost et al., 2015) and relies on a late fusion of the audio and visual modalities. This second approach focuses on the shot sequences described in Subsection 3.2.3 with two alternating and recurring shots, each one corresponding to one of the two speakers involved. Once automatically detected, such patterns limit the interaction scheme in which diarization is performed. We therefore perform independently audio and video speaker diarization, before merging the resulting partitions of the spoken segments in an optimal way. The two modalities are expected to be uncorrelated in their respective mistakes, and to compensate each other.

Both methods rely on the acoustic features described in Subsection 3.5.2.

The next, global speaker diarization step aims at detecting the recurring speakers from on scene to the other, and take as inputs the speakers locally hypothesized in each dialogue sequence during the local step, whatever the method used. This global clustering step is described in Subsection 3.5.5.

3.5.2 Acoustic features for local speaker diarization

Subtitles

Widely available, the subtitles of TV serial episodes are here used to approximate utterance boundaries. As accurate transcriptions of utterances, they usually match them temporally, despite some slight and unpredictable latency before they are displayed and after they disappear. When such a latency was too high, we manually adjusted the utterance boundaries.

Moreover, each subtitle is generally associated with a single speaker, and in the remaining cases where two speakers are involved in a single subtitle, speech turns are displayed onscreen, allowing to split the whole subtitle into the two corresponding utterances.

The detection of change points between the possible audio sources, as a prerequisite of most diarization systems, is therefore here avoided, allowing us to focus on the

clustering process.

I-vectors

The acoustic parameterization of utterances relies, as a state-of-the-art technique used in the speaker verification field, on the i-vectors model (Dehak et al., 2011).

After 19 cepstral coefficients plus energy are extracted, a 512-component GMM / UBM[6] is trained on the corpus subset (described in Subsection 3.5.6) we used for the speaker diarization task; the total variability matrix is then trained on all the spoken segments contained in the corpus subset[7], and 60-dimensional[8] normalized i-vectors are finally extracted, each associated with a single utterance. I-vectors are extracted using the ALIZE toolkit, described in (Larcher et al., 2013). The initial set of instances to be clustered therefore consists of speech segments parameterized by i-vectors.

3.5.3 Local speaker diarization (1) : mono-modal

Speaker diarization is first locally performed as an agglomerative clustering process within each dialogue sequence, as delimited by the use of the visual patterns described in Subsection 3.2.3 and captured by the extended version of the regular expression 3.1.

For the set of utterances $\mathbf{u}(l_1, l_2)$ covered by the pattern $(l_1, l_2) \in \mathcal{P}(\mathbf{s})$ (defined by Equation 3.2), the bottom-up clustering algorithm relies on the following:

- The Mahalanobis distance is chosen as a distance metrics between the i-vectors corresponding to the spoken segments, resulting in a matrix M of distances between the utterances contained in $\mathbf{u}(l_1, l_2)$.

  The covariance matrix used to compute the Mahalanobis distance is the within class covariance matrix of the training set, as described in (Bousquet et al., 2011), and is computed as follows:

$$W = \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n_s} (\mathbf{u_i^s} - \overline{\mathbf{u_s}})(\mathbf{u_i^s} - \overline{\mathbf{u_s}})^T \tag{3.5}$$

  where n denotes the number of spoken segments of the training set (described in Subsection 3.5.6), S the number of speakers and n_s the number of segments uttered by the speaker s; $\overline{\mathbf{u_s}}$ is the mean of the i-vectors corresponding to utterances of speaker s and $\mathbf{u_i^s}$ denotes the i-vector corresponding to the i^{th} utterance of speaker s.

- The Ward's aggregation criterion is used during the agglomeration process to estimate the distance $\Delta I(c, c')$ between the clusters c and c'; it is computed as follows:

[6] A 64-component GMM / UBM is used for the alternative multi-modal speaker diarization framework.

[7] The total variability matrix is only trained on the segments of the currently processed episode in the alternative multi-modal framework.

[8] 20-dimensional i-vectors in the alternative multi-modal speaker diarization framework.

$$\Delta I(c, c') = \frac{m_c m_{c'}}{m_c + m_{c'}} d^2(g_c, g_{c'}) \tag{3.6}$$

where m_c and $m_{c'}$ are the respective masses of the two clusters, g_c and $g_{c'}$ their respective mass centers and $d(g_c, g_{c'})$ the distance between the mass centers.

- Finally, the Silhouette method is used to cut the dendogram resulting of the clustering process and obtain the final partition of the spoken segments. Described in (Rousseeuw, 1987), the Silhouette method allows to automatically choose a convenient partition of the instance set by evaluating the quality of each possible partition resulting from the clustering process. For a given partition, if instances appear closer to another cluster than to their own, the quality measure tends to decrease, and to increase if the instances are appropriately assigned to their respective clusters.

3.5.4 Local speaker diarization (2) : multi-modal alternative

Alternatively, the local speaker diarization step, rather than relying on audio-only features, can be performed in a multi-modal way: the utterances contained in the dialogue sequences captured by the regular expression 3.1, besides acoustic i-vector based features, can also be described by visual features.

Visual representation of utterances

The visual parameterization of a spoken segment relies on its synchrony with the character being filmed while the segment is uttered: the utterance is parameterized according to its temporal distribution over the shots, as labeled according to their similarities (see Subsection 3.2.2). Considering the set $\Sigma = \{l_1, ..., l_m\}$ of shot labels involved in a movie, the i^{th} dimension of $\mathbb{R}_+^m$ is associated with the i^{th} shot label. Each utterance $\mathbf{u} = (u_1, ..., u_m)^\intercal$ is then described as an m-dimensional vector, where the i^{th} component u_i corresponds to the overlapping time in seconds between the utterance $\mathbf{u}$ and the shot label l_i. Figure 3.11 shows another example of a shot sequence matching the regular expression 3.1, with two alternating, recurring shots typical of dialogue sequences.

Figure 3.11 – *Example of shot sequence ...$l_1 l_2 l_1 l_2 l_1$... with two recurring shots, respectively labeled l_1 and l_2, captured by the regular expression 3.1.*

Figure 3.12 shows the distribution over time of the two alternating shots of Figure 3.11, here labeled (c_{126}, c_{127}). The top line reports the alternation of both shots over time; the bottom line contains the utterances covered by the sequence.

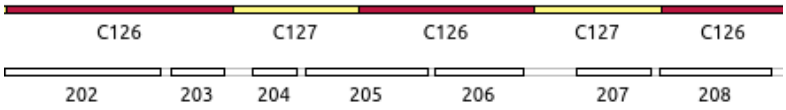


Figure 3.12 – *Shot sequence ...$c_{126}c_{127}c_{126}c_{127}c_{126}$... for two shot labels c_{126} and c_{127} (top line) with the covered utterances (bottom line).*

The utterance $\mathbf{u}^{(\mathbf{205})}$ for instance overlaps the two shots and is set to 1.56 (seconds) for its 126$^{\text{th}}$ component, to 1.16 for its 127$^{\text{th}}$ component, and to zero for all the other ones.

P-median clustering

The n utterances covered by a particular pattern can then be described either according to audio-only features, resulting in a set $\mathcal{U}_a$ of n 20-dimensional i-vectors, or by using visual-only features, resulting in a set $\mathcal{U}_v$ of n m-dimensional vectors, where m denotes the number of shot labels in the considered episode.

Both sets $\mathcal{U}_a$ and $\mathcal{U}_v$ are first partitioned into two clusters each: the average number of speakers in the dialogue patterns (1.77, reported in Table 3.3) leads us to perform such a bipartition of the instance set.

With such a fixed number of clusters, the partition problem can be modeled as a p-median problem. The p-median problem (Hakimi, 1964, 1965) belongs to the family of facility location problems: p facilities must be located among possible candidate sites such that the total distance between demand nodes and the nearest facility is minimized. The p-median problem can be transposed into the cluster analysis context (Klastorin, 1985) with a predefined number of classes. The instances to cluster into p classes correspond to the demand nodes and each instance may be chosen as one of the p class centers. Choosing the centers so as to minimize the total distance between the instances and their nearest center results in compact classes with medoid centers, defined as the most central instances in their class.

Considering the set $\mathcal{U}$ of n utterances covered by a pattern, the clustering problem can be modeled using the following binary decision variables: $x_i = 1$ if the i^{th} utterance $\mathbf{u}^{(\mathbf{i})}$ is selected as one of the p cluster centers, $x_i = 0$ otherwise; $y_{ij} = 1$ if $\mathbf{u}^{(\mathbf{i})}$ is assigned to the cluster center $\mathbf{u}^{(\mathbf{j})}$, $y_{ij} = 0$ otherwise. The model constants are the number of centers p and the distance coefficients d_{ij} between the utterances $\mathbf{u}^{(\mathbf{i})}$ and $\mathbf{u}^{(\mathbf{j})}$, measured by using the normalized euclidean distance between the corresponding vectors, either audio or video-based. The p-median clustering problem can then be modeled as the following integer linear program, closely related to the program described in (Dupuy et al., 2012, 2014) in the same speaker diarization context:

$$\text{(P1)} \begin{cases} \min \left(\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} y_{ij} \right) \\ \text{s.t.} \begin{cases} \sum_{j=1}^{n} y_{ij} = 1 & i = 1, ..., n \\ \sum_{i=1}^{n} x_i = p \\ y_{ij} \leqslant x_j & i = 1, ..., n; j = 1, ..., n \\ x_i \in \{0, 1\} & i = 1, ..., n \\ y_{ij} \in \{0, 1\} & i = 1, ..., n; j = 1, ..., n \end{cases} \end{cases} \tag{3.7}$$

The first constraints $\sum_{j=1}^{n} y_{ij} = 1$ $(i = 1, ..., n)$ ensure that each utterance is assigned to exactly one center; the second one $\sum_{i=1}^{n} x_i = p$ that exactly p centers are chosen; the third ones $y_{ij} \leqslant x_i$ $(i = 1, ..., n; j = 1, ..., n)$ prevent an utterance from being assigned to a non-center one ($y_{ij} = 0$ if $x_j = 0$).

Setting $p := 2$ and solving twice the integer linear program 3.7, once for the set $\mathcal{U}_a$ of utterances described by audio features, and then for the set $\mathcal{U}_v$ of utterances described by visual ones, results in two distinct bipartitions of the same utterance set $\mathcal{U}$.

Optimal matching fusion

The two bipartitions of the utterance set are then merged by solving the classical maximum weighted matching in a bipartite graph[9].

On Figure 3.13, the set of utterances $\mathcal{U} = \{\mathbf{u^{(1)}}, \mathbf{u^{(2)}}, \mathbf{u^{(3)}}, \mathbf{u^{(4)}}\}$ is twice partitioned using the audio and video modalities, resulting in two different partitions $\mathcal{Q}_a = \{Q_a^{(1)}, Q_a^{(2)}\}$ and $\mathcal{Q}_v = \{Q_v^{(1)}, Q_v^{(2)}\}$.

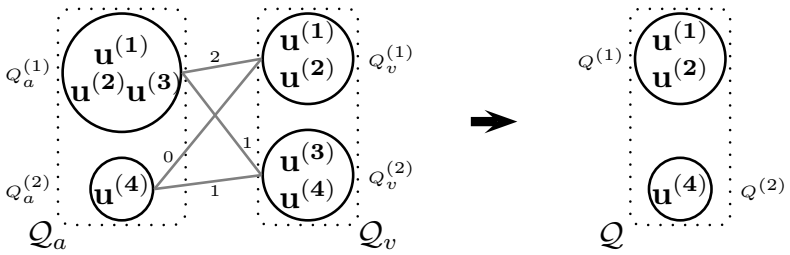


Figure 3.13 – *Fusion of two partitions by maximum weighted matching in a bipartite graph.*

A bipartite weighted graph $\mathcal{G} = (\mathcal{Q}_a, \mathcal{Q}_v, \mathcal{E})$, where $\mathcal{E} = \mathcal{Q}_a \times \mathcal{Q}_v$, can then be

[9] See for instance (Bunke et al., 2007) for using bipartite graph matching in a slightly different context, as a way to measure the distance between two partitions of the same set.

defined by assigning to each edge $(Q_a^{(i)}, Q_v^{(j)}) \in \mathcal{E}$ a weight w_{ij} corresponding to the sum of the duration of the utterances that the sets $Q_a^{(i)}$ and $Q_v^{(j)}$ have in common. In the example of Fig. 3.13, the edges of the bipartite graph are weighted assuming a same duration of 1 for all the utterances $\mathbf{u}^{(1)}, ..., \mathbf{u}^{(4)}$.

The best matching between both partitions consists in choosing non-adjacent edges (without any node in common) so that the sum of their weights is maximized. By using a decision variable y_{ij} such that $y_{ij} = 1$ if the edge $(Q_a^{(i)}, Q_v^{(j)})$ is chosen, $y_{ij} = 0$ otherwise, the general problem for two partitions respectively containing p and p' subsets can be modeled as follows:

$$\text{(P2)} \begin{cases} \max \left(\sum_{i=1}^{p} \sum_{j=1}^{p'} w_{ij} y_{ij} \right) \\ \text{s.t.} \begin{cases} \sum_{j=1}^{p'} y_{ij} \leqslant 1 & i = 1, ..., p \\ \sum_{i=1}^{p} y_{ij} \leqslant 1 & j = 1, ..., p' \\ y_{ij} \in \{0, 1\} \quad (i, j) \in \{1, ..., p\} \times \{1, ..., p'\} \end{cases} \end{cases} \tag{3.8}$$

The first and second constraints ensure that only non adjacent edges can be selected.

In the example of Fig. 3.13, where $p := p' := 2$, the optimal choice consists in assigning $Q_a^{(1)}$ to $Q_v^{(1)}$ and $Q_a^{(2)}$ to $Q_v^{(2)}$, for a total cost of 3.

It is well-known that the constraints matrix of the integer linear program 3.8 is totally unimodular (Nemhauser et Wolsey, 1988). Such a property ensures the integrity of the optimal solution of the continuous relaxation of the program 3.8, resulting in the following formulation, computationally much more straightforward to solve:

$$\text{(P3)} \begin{cases} \max \left(\sum_{i=1}^{p} \sum_{j=1}^{p'} w_{ij} y_{ij} \right) \\ \text{s.t.} \begin{cases} \sum_{j=1}^{p'} y_{ij} \leqslant 1 & i = 1, ..., p \\ \sum_{i=1}^{p} y_{ij} \leqslant 1 & j = 1, ..., p' \\ y_{ij} \in [0, 1] \quad (i, j) \in \{1, ..., p\} \times \{1, ..., p'\} \end{cases} \end{cases} \tag{3.9}$$

where the integrity constraint on the variables y_{ij} has been relaxed.

Once the matching choice is made by solving the program 3.9, we keep the intersection of the matching subsets, resulting in a new set $\mathcal{Q}$ of subsets of $\mathcal{U}$ corresponding to

cases of agreement between the two modalities: each of the obtained subsets is expected to contain segments both acoustically close to each other and uttered as the corresponding speaker is being filmed. Conversely, the residual segments ($\mathbf{u}^{(3)}$ in the example of Fig. 3.13) are discarded as cases of disagreement between the audio and visual modalities, either because the utterance is acoustically ambiguous or because of asynchrony between the utterance and the character being currently filmed.

Re-assignment of discarded utterances

We finally re-assign the residual utterances to the closest medoid of the refined clusters resulting from the combination of the audio and visual modalities. This stage of re-assignment relies on the audio-only features of the remaining utterances: possibly discarded because of their visual asynchrony, such utterances might not be correctly re-assigned by relying on the visual modality. On the other hand, using the audio modality to achieve such a re-assignment is expected to be more robust than when performing the audio-only clustering described above: by using medoids of clusters refined by the use of the video modality, some errors made during the audio-only stage are expected to be here avoided. Moreover, the medoid, being less sensitive to outliers than the centroid, is expected to properly handle the case of impure clusters containing isolated misclassified utterances resulting from a joint error of both modalities. Medoids are retrieved by solving again for each cluster a 1-median problem.

3.5.5 Constrained global clustering

Once speaker diarization is performed within each dialogue sequence, we perform a second stage of clustering in order to retrieve the recurring speakers.

We first concatenate into a single audio file the sequence of all segments locally hypothesized as uttered by the same i^{th} speaker and model it by a normalized speaker i-vector $\mathbf{s_i}$ of 60 components. The global clustering of the resulting speaker i-vectors is performed in the same way than the local mono-modal speaker diarization step, by using Mahalanobis distance based on the W covariance matrix as computed from Equation 3.5, Ward's aggregation criterion 3.6 and the Silhouette method to extract the final partition of speakers. However, this second step is guided, at each agglomeration step, by the structural information given by the visual segmentation of the movie into dialogue sequences as described in Subsection 3.2.3: the global clustering step has to prevent speakers locally hypothesized to be distinct from being assigned to the same cluster during the iterative agglomeration process. The integration of such a cannot-link constraint in the bottom-up clustering algorithm is achieved in the following way:

- In the initial matrix M of the distances between the i-vectors corresponding to the locally hypothesized speakers, the distance $d(\mathbf{s}, \mathbf{s}')$ between two instances $\mathbf{s}$ and $\mathbf{s}'$ is set to $+\infty$ if the corresponding two speakers appear together in the same dialogue scene:

$$d(\mathbf{s}, \mathbf{s}') = +\infty \iff \exists (l_1, l_2),\ \mathbf{u}(\mathbf{s}) \cup \mathbf{u}(\mathbf{s}') \subseteq \mathbf{u}(l_1, l_2) \tag{3.10}$$

where (l_1, l_2) denotes a dialogue pattern, $\mathbf{u}(l_1, l_2)$ the set of utterances covered by the pattern (l_1, l_2), and $\mathbf{u}(\mathbf{s})$ the set of utterances assigned to the speaker $\mathbf{s}$ during the local clustering step.

- The distance $\Delta I(c, c')$ between the clusters c and c' is set to $+\infty$ if at least one instance of the first cluster is located at an infinite distance from an instance of the second one:

$$\Delta I(c, c') = +\infty \iff \exists (\mathbf{s}, \mathbf{s}') \in c \times c',\ d(\mathbf{s}, \mathbf{s}') = +\infty \tag{3.11}$$

where $\mathbf{s}$ and $\mathbf{s}'$ denote i-vectors corresponding to hypothesized speakers.

The application of Rules 3.10 and 3.11 prevents two distinct speakers from being clustered into the same subset when choosing at each iteration of the agglomerative process the two closest instances to merge. Fig. 3.14 illustrates both the application of these rules at the initial step of the agglomerative process and how this "different-speakers" property is inherited by the newly created cluster. Local dialogue sequences are surrounded by dotted rectangles; each node s_{ij} represents the i^{th} speaker hypothesized in the j^{th} dialogue sequence; the edges between two nodes represent their distance; the absence of edge between two nodes corresponds to an infinite distance. Merging the two closest nodes s_{11} and s_{12} results in an isolated cluster $s_{11}s_{12}$ that both inherits from the difference between the two speakers of the first scene and from the difference between those of the second one: the hypothesized recurring speaker in the two scenes has indeed to be different from the speakers he is respectively talking to within both scenes.

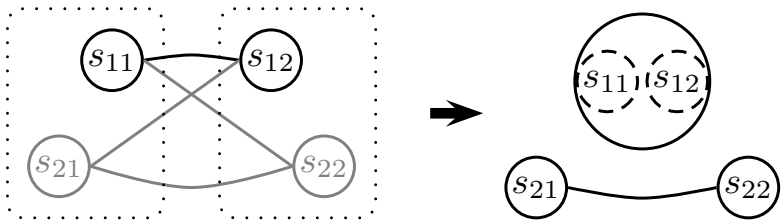


Figure 3.14 – *First iteration of constrained global clustering.*

Such a "different-speakers" property, as propagated at each step of the agglomerative process, is expected to prevent the speakers involved in a same dialogue to be prematurely clustered: the background music of a dialogue may for instance hide the inter-speaker variability and cause such an early clustering.

Moreover, the main consequence of respecting such a constraint is to block the clustering process before assigning all the instances to the same cluster. In the small example of Figure 3.14, only one more step of the agglomerative process could be achieved, by clustering s_{21} and s_{22}: the narrative structure (two dialogues with two speakers each) remains indeed compatible with such a clustering. The resulting dendogram would then be split into two distinct trees, resulting in an irreducible partition (Davidson et

Ravi, 2009). Fig. 3.15 shows dendograms corresponding to agglomerative clustering of local speakers, as hypothesized in one episode of *Breaking Bad*. The one figuring on top is obtained in a classical way, but may be difficult to cut automatically to extract the best partition of the instance set. The bottom part of the figure, obtained with the same data by integrating the "different-speakers" property to the clustering process, shows five trees corresponding to five incompatible groups of speakers; each one is made of a group of narratively consistent speakers, with possibly several occurrences of the same one.

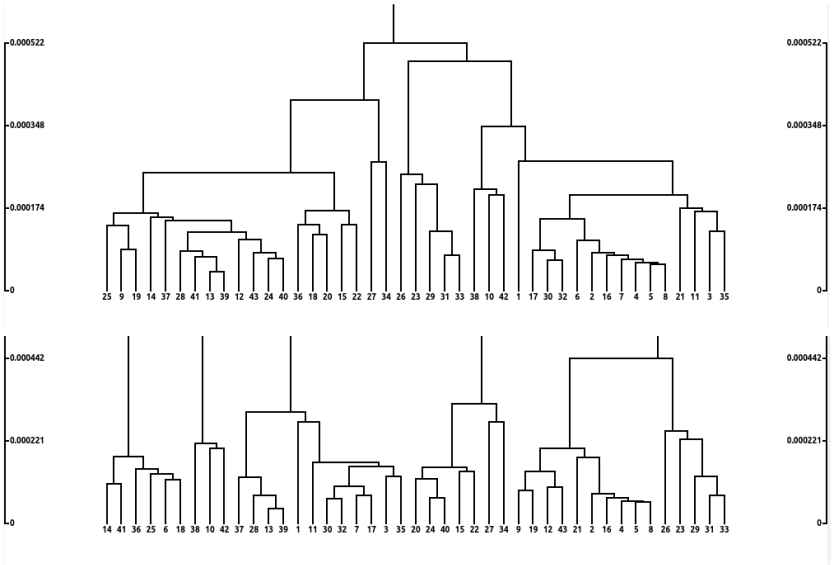


Figure 3.15 – *Dendograms obtained by agglomerative clustering on local speaker hypotheses, unconstrained (top); constrained (bottom).*

Each of these remaining trees of compatible speakers is finally cut using the Silhouette method described in Subsection 3.5.3 and the final partition of the instance set is obtained by the union of the partitions obtained for each tree.

However, this constrained global clustering step remains dependent on the outputs of the local one. If a single speaker is wrongly split into two clusters during the local clustering step, the two resulting utterance groups will never be merged during a global clustering embedding the "different-speakers" property. Nevertheless, even during an unconstrained clustering process, such groups would be merged lately, possibly after the best partition is reached.

3.5.6 Experiments and results

Corpus

The corpus subset we used for performing speaker diarization consists of the same 9 episodes as the ones used in Section 3.2: the first three episodes of *Breaking Bad*, *Game*

of Thrones, and *House of Cards*. In every episode, we manually labeled every subtitle according to the corresponding speaker, both for training and evaluation purposes. The total speech duration in these nine episodes amounts to a bit more than three hours (3:12). A subset of six episodes (denoted DEV) was used for development purpose when needed, mainly for the computation of the covariance matrix from Equation 3.5; the remaining three ones (denoted TEST) were used for test purpose.

We first describe the experimental results we obtain when locally performing speaker diarization within each scene visually hypothesized as a dialogue: this local speaker diarization step comes in two alternative ways, either mono, or multi-modal.

Local speaker diarization (1) : mono-modal

The Diarization Error Rate (DER) used to evaluate the local clustering step is computed independently in each episode dialogue before averaging the obtained scores according to each dialogue duration: this evaluation metrics is denoted as the *single-show* DER in (Rouvier et al., 2013). The results are reported in Table 3.7, when using both the reference (denoted *input ref.*) and the automatically detected (denoted *input auto.*) similar shots. For the sake of comparison, agglomerative clustering (denoted AC), is compared to a "naive method", very close to the visual clustering of utterances we described in Subsection 3.5.4. This approach relies on a strong assumption of synchronization between the audio and video streams: the local utterances are clustered by assigning to each speech segment the label of the current shot, assuming the two alternating shots match exactly the speaker turns.

***Table 3.7** – Single-show DER by episode obtained for the local diarization step*

TV serial	episode	**input auto.**		**input ref.**	
		naive	AC	*naive*	AC
BB	S01E01	30.26	**19.11**	22.81	21.00
	S01E02	22.06	22.51	19.78	**19.14**
GoT	S01E01	22.16	23.70	19.46	**15.78**
	S01E02	26.19	18.78	22.80	**16.61**
HoC	S01E01	17.23	13.36	16.31	**11.84**
	S01E02	30.66	**18.18**	31.87	19.12
average DEV		24.76	19.27	22.17	**17.25**
BB	S01E03	40.45	21.15	24.31	**12.15**
GoT	S01E03	33.45	17.43	35.43	**12.80**
HoC	S01E03	24.44	12.83	22.95	**12.82**
average TEST		32.78	17.14	27.56	**12.59**

As can be seen, the results obtained by performing an audio-based clustering of the utterances of each dialogue sequence appear better than those obtained by applying the naive image-based method, possibly mistaken by audiovisual asynchrony.

Furthermore, the results obtained when performing agglomerative clustering on the test set turn out to be even better than those obtained on the development set: both

the total variability and within-class covariance matrices described in Subsection 3.5.2 seem to generalize well.

Finally, the automation of the previous shot processing step, though degrading performance on the development and test sets, does not impact it significantly, which confirms the reliability of the visual modality.

Local speaker diarization (2) : multi-modal alternative

The single-show DER is also used when evaluating the alternative, multi-modal approach we introduced in Subsection 3.5.4 for locally performing speaker diarization within each dialogue sequence. Results are first given when using a single modality, either audio or video. The optimal matching (denoted *om*) performed during the multi-modal fusion step is then evaluated in two ways: first by discarding from scoring the utterances for which the two modalities disagree (denoted *om-ra*, for *optimal matching with no re-assignment step*). In this case, the resulting speech coverage of the scored utterances is indicated in % in parenthesis. Moreover, results are also given when the optimal matching between both modalities is followed by a step of audio reassignment of the remaining utterances (denoted *om+ra*). For the sake of comparison, the results obtained by optimizing jointly in a weighted sum (denoted *ws*) the two *p*-median mono-modal objective functions are also reported. Finally, an oracle score is estimated by manually labeling the utterances according to the reference speaker when at least one of both modalities succeeds in retrieving it.

***Table 3.8** – Single show DER obtained for all episodes.*

TV serial	episode	**mono-modal**		**oracle**	**multi-modal**		
		audio	*video*		*om-ra*	*om+ra*	*ws*
BB	S01E01	25.2	26.9	8.3	18.0 (67.7)	24.0	26.9
	S01E02	26.6	24.5	8.2	17.2 (69.7)	20.4	24.5
	S01E03	26.8	26.9	9.6	17.1 (67.4)	24.7	27.3
GoT	S01E01	22.6	24.7	7.6	13.1 (69.2)	21.1	24.5
	S01E02	28.7	27.7	10.2	20.0 (68.2)	25.9	27.0
	S01E03	12.8	29.4	5.3	9.9 (71.1)	13.2	28.2
HoC	S01E01	17.5	21.9	3.8	10.0 (71.6)	17.7	22.2
	S01E02	21.4	29.4	10.2	15.4 (70.6)	20.8	29.4
	S01E03	20.6	25.6	6.9	12.8 (70.2)	20.6	25.4
average		**22.5**	**26.3**	**7.8**	**14.8** (69.5)	**20.9**	26.2

As can be seen, the results obtained by performing mono-modal speaker diarization are in average better for the audio modality than for the video one[10].

[10]The error rates obtained when clustering the utterances from audio-only features turn out to be higher than the results described above for the mono-modal approach: both approaches are indeed quite different and the results are not comparable. The mono-modal approach described in Subsection 3.5.3, relying on a covariance matrix computed (Equation 3.5) from annotated data, is partially supervised, whereas the *p*-median clustering evaluated here is fully unsupervised. Moreover, the output of the preliminary image processing subtasks are here fully automatic. Finally, the dialogue shot patterns used here target two-

Nonetheless, the computed oracle shows that both modalities are not redundant: by managing to combine them perfectly (*oracle*), the DER would decrease dramatically (from 22.5% to 7.8% for the audio modality, and from 26.3% to 7.8% for the video one), which confirms that both these modalities are highly complementary for the speaker diarization task and that the errors made are not correlated. Indeed, there is no reason for a speaker to talk in an atypical way, possibly resulting in acoustic misclustering, whenever he/she is not appearing onscreen (audiovisual asynchrony resulting in visual misclustering).

Moreover, when both modalities are combined (*om-ra*), resulting in a new partial clustering of the utterance set, the DER remains relatively low if about 30% of the utterances, corresponding to cases of disagreement between both modalities, are discarded from the evaluation (DER amounting to 14.8% for 69.5% of speech covered).

As expected, while processing the critical 30% remaining utterances (*om+ra*), the DER tends to increase (from 14.8% to 20.9% in average) but is still lower than the DER obtained when combining in a weighted sum (*ws*) both modalities and when using the single audio features (22.5%): compared to the audio-only modality, the relative improvement amounts to 7.11% in average.

Furthermore, the results obtained when combining both modalities are equal or better than when using a single one for 7 episodes out of 9; and even in the remaining two cases (GoT, S01E03 and HoC, S01E01), they turn out to be very close to the best modality.

Constrained global speaker diarization

Table 3.9 – DER *obtained for the global diarization step.*

TV serial	episode	**input auto.**		**input ref.**		**spch ref.**	
		2S	*cst. 2S*	*2S*	*cst. 2S*	LIA	LIUM
BB	S01E01	51.36	56.00	52.66	**48.10**	72.06	67.21
	S01E02	**41.83**	65.07	58.76	49.49	77.03	76.79
GoT	S01E01	70.13	**52.79**	70.67	53.87	65.57	58.49
	S01E02	67.28	**38.85**	70.32	41.24	65.29	60.80
HoC	S01E01	**50.04**	55.61	52.70	52.15	60.26	62.37
	S01E02	64.91	56.40	63.65	**37.09**	67.05	59.00
average DEV		57.59	54.11	61.46	**46.99**	67.88	64.11
BB	S01E03	60.41	**33.94**	59.22	42.64	60.61	55.56
GoT	S01E03	74.71	**49.31**	70.34	63.17	61.33	52.89
HoC	S01E03	**57.68**	59.87	67.52	67.41	70.55	67.05
average TEST		64.13	**47.71**	65.69	57.74	64.16	58.50

Table 3.9 reports the results obtained when clustering the local speakers as output by applying the mono-modal approach described in Subsection 3.5.3, completing the sec-

character dialogues and rely on the basic version of the regular expression 3.1, instead of the extended one.

ond step of the speaker diarization process. Results are given both when taking as input the local speakers hypothesized in each dialogue sequence during the local step (*input auto.*) as well as the real speakers manually annotated (denoted *input ref.*). In both cases, the second step of clustering is performed in an unconstrained way (denoted *2S*), allowing any local speakers to be clustered during the agglomerative process, and in a constrained way (denoted *cst. 2S*), by preventing it. For the sake of comparison, the results of two standard speaker diarization tools (denoted LIA, described in (Bozonnet et al., 2010), and LIUM, described in (Meignier et Merlin, 2010) and (Rouvier et al., 2013), are also reported: these tools receive as input all the spoken segments covered by the dialogue patterns.

Though still high, the DER is generally reduced by integrating to the clustering process the structural information based on visual patterns. By blocking the clustering before all the instances can be gathered, the "different-speakers" property allows to cut the resulting dendogram at a suitable level, providing the agglomerating process with an early stop condition, when only a few mutually exclusive groups of instances remain. By contrast, unconstrained clustering has to face the critical issue of finding the optimal partition of the instance set.

Table 3.10 reports the average number of speakers involved in the considered dialogue sequences of the different episodes, as hypothesized by the different systems.

Table 3.10 – DER *Average number of hypothesized speakers*

	truth	*2S*	*cst. 2S*	LIA	LIUM
bb	**10.3**	7.3	**11**	6	25.7
got	**25.3**	4.7	15.7	9.3	**24**
hoc	**20.7**	3.7	**24**	6	27

As can be seen, two systems (unconstrained 2-step clustering and LIA), tend to cut the clustering dendogram at a high level, resulting in a few number of too wide classes. Conversely, LIUM, by cutting the tree at a low level, overestimates in two cases the number of speakers. The constrained clustering approach (*cst. 2S*), resulting in disjoint dendograms, offers an reasonable approximation of the number of speakers and prevents early as well as late cuts of the clustering tree.

3.5.7 Speaker diarization: conclusion

To summarize, we proposed to perform speaker diarization within TV serial episodes by relying on the structural information they carry. As described in Subsection 3.2.3, shot patterns typical of dialogue sequences are first extracted and a preliminary step of speaker diarization can be locally performed within each dialogue scene, either from audio-only features, or from both acoustic and visual features.

The multi-modal alternative we introduced to locally perform speaker diarization within each dialogue consists in a late fusion of acoustic and visual features. Speaker

diarization is first performed separately for audio and visual features of the utterances by using the p-median model, before both resulting bipartitions of the utterance set are optimally matched into refined clusters corresponding to cases of agreement between both modalities. The remaining utterances for which both modalities disagree are then acoustically assigned to the closest medoid of the newly created clusters, expected to be more robust than those based on an audio-only approach. The experimental results obtained by using both modalities turn out to outperform those obtained by purely mono-modal approaches.

Once locally hypothesized within each dialogue sequence, the possibly recurring speakers are retrieved by performing a second step of clustering: at each iteration of this global clustering step, the constraint that speakers locally assumed to be different must not be clustered is propagated; as a result, the agglomerative process is blocked far before all the instances are clustered, allowing a more convenient partition of the initial set than when applying an unconstrained approach.

Though this two-step framework results in error rates significantly lower than when using audio-only standard speaker diarization tools, they remain too high to serve as a reliable basis for further modeling the narrative content of TV serials. Furthermore, the methods we introduced were evaluated on subsets of single episodes only, whereas our summaries rely on modeling the plot content over full seasons of TV serials, possibly containing several dozens of episodes involving hundreds of different speakers. At such scales, the results obtained when automatically performing speaker diarization would be even noisier. Finally, we wanted to do a relatively large scale study to evaluate our summaries and could not afford to show the viewers in a limited time both fully and partially automatic summaries to measure the impact of the errors made at the speaker diarization level. We shall therefore use hereafter the reference speakers, as manually annotated, both when further modeling the TV serial content (Chapter 4), and when generating the summaries (Chapter 5).

3.6 Conclusion

In this chapter, we described the various subtasks our summarization framework relies on. First, we detailed the video segmentation techniques we need to split the video stream into different levels of granularity: video shots, dialogue sequences matching predefined shot patterns, logical story units, and scenes. Logical story units of about 10 seconds are natural units for further insertion into summaries of TV serials: defined as possibly intertwined sequences of recurring shots, they allow to capture semantically homogeneous sequences, especially within dialogues between characters. We then detailed the mid-level features, along with the related extraction techniques, we will use in Chapter 5 when estimating the relevance of each LSU for possible selection and insertion into the summaries: shot size and background music are expected to capture stylistically salient sequences, whereas speaker labels, as resulting from speaker diarization, are more content-related and can be used, as detailed in the next chapter, as a possible basis for modeling the plot of TV serials from character interactions.

Chapter 4

Plot modeling: conversational network of characters

Contents

4.1 Introduction

Effective summaries of TV serials in the real-case scenario we described in Chapter 1 should both emotionally re-engage the viewers into the plot and provide them with a comprehensive recap of the narrative content of the past seasons.

There is a limited amount of works that attempt to automatically model the plot of a movie. In (Guha et al., 2015), the authors aim at automatically detecting from film grammar the typical three-act narrative structure of Hollywood full-length movies. Shot frequency, motion activity, along with musicality and speech rate, are successfully used as discriminant features for segmenting the whole stream into the usual three acts of *exposition*, *conflict*, and *resolution*. Nonetheless, this approach is purely based on stylistic patterns and does not provide any insight into the story content. Furthermore, the authors focus on a predefined narrative structure that generalizes with difficulty to the complex plots of TV serials.

As described in Chapter 2, *social network analysis* (SNA) has recently been applied to fictional networks of interacting characters, mostly extracted from literary works. Some authors deal with the pre-processing steps needed for identifying the characters involved in the plot – a much trickier task for novels than for theatre or movie screenplays. Once extracted, the social network of interacting characters is able to unveil underlying structures and to provide renewed views on classical literary works.

SNA-based approaches are more recent and sparse in the context of multimedia works. As mentioned in Chapter 2, (Weng et al., 2007, 2009) presented a method to automatically analyze the plot of a movie, and (Ercolessi et al., 2012) adopted a similar approach to cluster the scenes of two TV series into separate storylines. However, such works focus either on full length-movies or on standalone episodes of classical TV series, where character interactions are often well-structured into stable communities. Their approaches consequently do not necessarily translate well when applied to the continuous, possibly complex, plots of TV serials.

In this chapter, we introduce an SNA-based method aiming at automatically providing some insight into the complex plots of TV serials, while solving the limitations of the previous works. For this purpose, we do consider not only standalone episodes or full-length movies with stable and well-defined communities, but the complex plots of TV serials, as they evolve over dozens of episodes. In this case, no prior assumption can be made about a stable, static structure that would remain unchanged in every episode and that the story would only uncover, and we have to deal with evolving relationships, possibly temporarily linked into dynamic communities. In this case, we are left with building the current state of the relationships upon the story itself, which, by focusing alternatively on different characters in successive scenes, prevents us from monitoring instantaneously the full social network underlying the plot. We thus propose to address this problem by smoothing the sequentiality of the narrative, resulting in an instantaneous monitoring of the current state of any relation at some point of the story.

Our main contributions are the following:

- The first is *narrative smoothing*, the method we propose for the extraction of dynamic social networks of characters.
- The second is the experimental evaluation of the basic rules we introduce to estimate the interacting characters from the sequence of speech turns, once manually labeled according to the corresponding speakers.
- The third is a qualitative evaluation of our framework on these data, and a comparison with existing methods.

The rest of the chapter is organized as follows. In Section 4.2, we review in further details the previous works related to SNA-based plot identification. We then describe the method we propose, by first focusing in Section 4.3 on the way the verbal interactions between characters are estimated, before detailing in Section 4.4 the way a dynamic view of the relationships in TV serial plots can be built independently from the narrative pace. In Section 4.5, we first systematically evaluate the algorithm we use for estimating verbal interactions; then, we illustrate how our tool can be used by applying it to the three TV serials of our corpus, and we compare the obtained results to existing methods.

4.2 Previous works

In our review, we distinguish between two kinds of works: the first ones consider a static network resulting from the temporal integration over the whole considered period, which we call *complete aggregation*; the second ones extract and study a dynamic network based on a sequence of smaller integration periods called *time-slices*.

4.2.1 Complete aggregation

Cumulative networks were widely used when attempting to apply SNA for analyzing the plot of fictional works. The interactions are iteratively inserted as edges in the network of characters. They are possibly weighted and even directed, resulting in a static graph agglomerating all past relationships, whatever their time ordering.

In (Moretti, 2011), the author emphasizes and illustrates the role SNA can play to investigate the plot of literary works. First, after building the network of conversational interactions in a play, the plot, as a sequence of acts occurring over time, is frozen in a spatial, static view that exhibits some underlying patterns: for instance, the conversational network of verbal interactions in Shakespeare's *Hamlet* unveil some critical regions, such as the "region of death" in which the whole tragedy consists. Furthermore, a network-based definition of the protagonists should prevent scholars from applying binary, simplifying categories when considering the main and secondary characters. Finally, the SNA-based notion of community allows to exhibit two distinct spaces in *Hamlet*'s network: a space of legitimacy around Horatio, associated to the modern

democratic state, and a space of usurpation around Claudius, related to the old, declining monarchy. The author further illustrates the benefit of SNA for literary studies by considering the question of symmetry in Western and Chinese novels, both at the stylistic level and in the social network of interacting characters.

In (Weng et al., 2007, 2009), relying on similar observations, the authors make use of SNA to automatically analyze the plot of a movie. The social network of characters (denoted "RoleNet") is built as follows. They first manually characterize the scenes by their boundaries and the characters they involve. They then hypothesize an interaction between two characters whenever they both appear within the same scene. The network is obtained by representing characters as nodes and their interactions by links. These links are weighted according to the number of scenes in which they co-appear, resulting in a *cumulative* representation of time. The authors analyze this network through community detection. They apply this approach to so-called "bilateral movies", which involve only two major characters, each of them central in his own community. In (Weng et al., 2007), the *RoleNet* is used for further investigating the plot, by classifying scenes into one of the two storylines constituting a bilateral movie. In (Weng et al., 2009), an extended version of the network, without any prior assumption about the number of communities involved, is used as a basis for automatically detecting breakpoints in the story: a narrative breakpoint is assumed if the characters involved in successive scenes are socially distant in the network of characters, as accumulated over the whole story.

In (Ercolessi et al., 2012), a similar network of interacting speakers is used, among other features, for clustering scenes of two TV series episodes into separate storylines, defined as homogeneous narrative sequences related to major characters. A standard community detection algorithm is applied to the network of speakers, as built upon each episode, before the social similarity between any pair of scenes is computed, as a relevant high-level feature for clustering scenes into substories.

In summary, cumulative networks can be used as a reliable basis for automatically or manually analyzing the plot of fictional works with well-defined communities, as in plays, full-length movies or standalone episodes of classical TV series[1].

Nevertheless, for TV serials with complex, evolving and possibly parallel storylines, such a static approach is not appropriate. Indeed, a cumulative network built over a long period of time, as in TV serials, gets relatively dense and does not enable to extract meaningful information. More specifically, communities in the final agglomerative network undoubtedly always correspond to substories, partially disconnected in the narrative, but the opposite does not generally stand. Some individuals may have been strongly connected to each other at some point of the story, before some of them interact with other people for some time, resulting in a second substory. Once agglomerated in the cumulative network, such changes in the interaction patterns may be obscured. In some extreme cases, distinct narrative sequences may even result in a complete cumulative graph, for instance in the interaction pattern that follows:

[1]The website http://moviegalaxies.com/ (Kaminski et al., 2012) provides a convenient way of interactively visualizing such cumulative character networks for a database of about 700 movies.

$$s_{12}^{(1)} ... s_{12}^{(2)} s_{13}^{(3)} ... s_{13}^{(4)} s_{23}^{(5)} ... s_{23}^{(6)}$$

where $s_{ij}^{(t)}$ denotes the fact that the i^{th} and j^{th} characters are the only interacting speakers in the t^{th} episode. The three consecutive interaction sequences result in a triangular interaction pattern unable to reflect the three corresponding substories.

4.2.2 Time-slices

Some works attempt to take into account the evolution of the social network of the characters when analyzing the plot of fictional works. In (Agarwal et al., 2012), the authors emphasize the limitations of the static, cumulative graph when analyzing the centrality of the various characters of the novel *Alice in Wonderland*. A dynamic view of the social network is then introduced, by building successive static views of the network in every chapter, before standard centrality measures are separately computed in each of them and traced over time for some major characters. Each view corresponds to a so-called *time-slice*, or *time-window*.

Though widely used (Holme et Saramäki, 2012) when considering the evolution over time of general networks (*i. e.* not necessarily narrative ones), time-slice networks, as resulting from the differentiation over some time step of the cumulative network, may still be problematic. In (Clauset et Eagle, 2012), the authors focus on the critical issue of the time-slice duration, called "snapshot rate". It must be chosen carefully to allow to capture a sufficient amount of interactions, but not too many, otherwise one may obtain irrelevant network statistics. The authors then describe a way of automatically estimating the natural time-slice for monitoring over time the evolution of a network of daily contacts in a professional context, where the appropriate time-slice is expected to remain constant.

As a smoother alternative to fixed time-windowing, (Mutton, 2004) applies temporal decay to past interactions to estimate the current state of relationships between users of *internet relay chats*. The method is then extended to Shakespeare's plays for monitoring over time the evolution of the network of interacting characters. Nonetheless, we are left with the same kind of issue as with time-windowing approaches: as for the time-slice duration, the ideal value for the decay parameter may be tricky to set.

In order to model the plot of TV serials and allow further analysis, the time-slice should be short enough to capture punctual narrative events related to the social network of characters, but long enough to provide a comprehensive view of the state of the relationships at any point in the story. Unfortunately, getting such a snapshot of the current state of the relationships between the protagonists turns out to be particularly challenging. Unlike the network of physical contacts described in (Clauset et Eagle, 2012), the state of the relationships within a story is not fully monitored at any moment, but has to be inferred from the story itself. The narrative usually focuses alternatively on some relationships, possibly belonging to parallel storylines, and only provides a partial view on the network's current state. Some relationships may even

take place at the same moment in different places, but will be shown sequentially in successive scenes. Fig. 4.1 illustrates the typical sequential nature of the story as being narrated: three disjoint sets of interacting speakers, possibly at the same time but in different places, are shown sequentially in the story in three successive scenes.

Figure 4.1 – *Three different sets of interacting characters from three consecutive scenes.*

As a consequence, the temporalness of the narrative may be quite different from the temporalness of the underlying network: in particular, the only fact that a group of mutually interacting characters temporarily disappears from the story does not imply that the corresponding characters disappeared from the network. The narrative focus on these interacting characters may only have been postponed by the filmmaker. Furthermore, the pace of activation of the relationships in the story remains largely unpredictable, especially when multiple, disjoint storylines take place in parallel within the narrative. Fig. 4.2 plots the scene occurrences of three major character-based storylines in the first two seasons of *Game of Thrones*. Except in the very beginning of the first season, where Jon Snow and Tyrion Lannister meet each other, the three characters interact within well-separated communities.

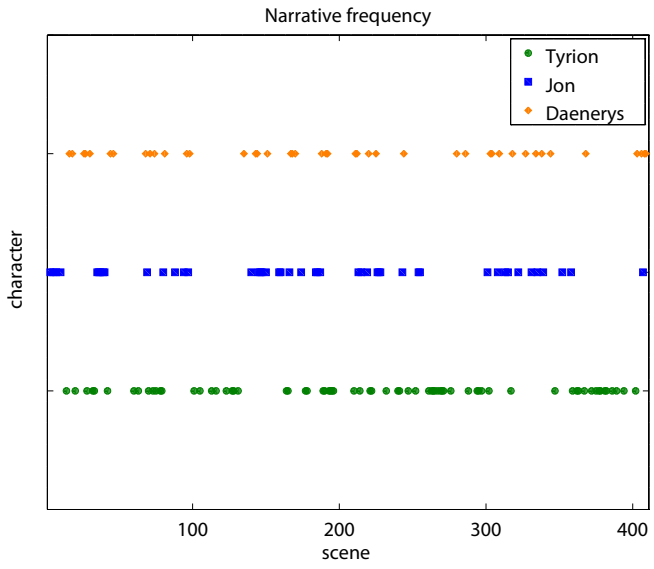


Figure 4.2 – *Narrative frequency of three character-based storylines in the first two seasons of Game of Thrones.*

As can be seen, the way the story alternatively activates these three major storylines does not seem to follow regular patterns. In such a case, the "ideal" time-slice may be

tricky to set. If too large, it will possibly mask the fast changes usually occurring in the most frequently activated storyline, for instance the story centered around Tyrion. If too narrow, it would lead to irrelevant interpretations of the narrative disappearance of some groups of relationships: the narrative disappearance of Jon Snow's storyline from scene 250 up to scene 300 does definitely not imply that he does not remain socially active in the meantime in his own community. Therefore, the sequential nature of the story should prevent us from identifying the time of the narrative to the time objectively affecting the social network that the story sequentially unveils.

In this chapter, we introduce a novel way of building the dynamic network of interactions between the characters of TV serials that allows to fully capture the instantaneous state of every relationship at any point of the story, whatever the pace of activation of each storyline in the narrative. The framework we introduce will prove in Chapter 5 to be especially well-suited to further modeling and summarizing each character's storyline. In the following two sections, we detail the two steps constituting our method: we first explain how we identify and characterize interactions between TV serial characters; then, we describe how we extract a smoothed dynamic network from the set of interactions, as hypothesized during the previous step.

4.3 From speaker diarization to verbal interactions

Getting an accurate view of verbal interactions within TV serials turns out to be quite challenging, either manually or automatically. When considering a sequence of speech turns within a scene, verbal relationships can be stated as soon as a speaker is talking to an audience, resulting in a directed conversational network, depending on whether someone is talking to, or is being talked by someone. But when a recorded conversation involves more than two speakers, stating who is talking to whom may be tricky. The sequence of speech turns does not convey in itself such an information. By just considering such a sequence, as the excerpt of *House of cards* shown on top of Fig. 4.3, it is impossible to guess who of the three involved speakers is actually speaking to whom.

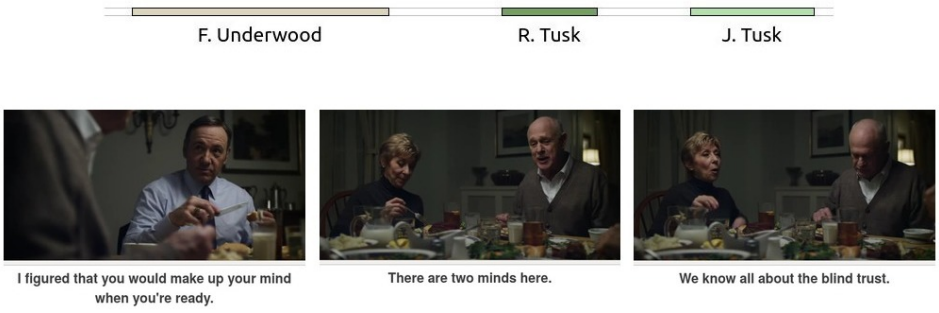


Figure 4.3 – *Sequence of speech segments (top) along with corresponding video frames and subtitles (bottom)*

Such an information is usually inferred from complementary sources when available, such as the semantic content of the utterances and/or the video recording of the conversation. On Fig. 4.3 the images corresponding to the three utterances, along with their linguistic content, help to disambiguate the verbal situation: from who they are looking at when speaking and/or from the use of personal pronouns, we infer that the first speaker (F. Underwood) is clearly speaking to the second one (R. Tusk), while the third one (J. Tusk) is speaking to the first one. However, guessing to whom the second character is specifically talking remains tricky, even from a careful human expertise using this additional information.

In order to avoid such a tedious human expertise, two main options are here considered for automatically estimating verbal interactions: the first one relies on the co-occurrence of speakers in scenes, while the second one relies on the sequence of utterances within each scene.

4.3.1 Scene co-occurrence

Verbal relationships between characters can first be indirectly deduced from their co-appearance within semantically homogeneous units. For TV serials, which tend to develop complex storylines over several seasons, each typically consisting of a dozen of episodes, possible units are seasons, episodes, or scenes. Seasons or even episodes turn out to be too wide units to provide an accurate view of the actual verbal interactions within TV serials. Because of parallel storylines, stating as interlocutors all the characters co-occurring in a single season or even episode would result in many irrelevant interactions. Considering the scene as a unit is much wiser: as stated in Subsection 3.2.5, a scene in a movie is defined as a homogeneous sequence of actions occurring at the same place within a continuous period of time. The characters co-appearing in a single scene are therefore expected to speak with one another.

A second choice has to be made concerning the kind of character appearance within scenes. Many scenes in movies contain passive characters who do not play any role in the plot and are only physically present but may be talked to by others. Verbal involvement turns out to be much more significant. Although non-verbal relationships, denoted as "observations" in (Agarwal et al., 2013), are still possible between main characters, for instance by only thinking of or by looking at someone else, they usually end up showing verbally in movies. So, by "characters", we will always mean "speakers", and by "occurrence" of the character within a scene, we will mean verbal involvement.

When speech turns, as well as scene boundaries, are explicit, as in plays or movie scripts, the verbal interactions estimated from speaker co-occurrence can be deduced in a fully automatic way. But when the play or the movie is only available as a recorded performance, this information has to be retrieved, either automatically or manually. For TV serials in particular, the scripts are not easily available, or contain only unnormalized and partial information provided on the Web by communities of viewers: we are then left with retrieving speakers as well as scene boundaries.

As stated in Subsection 3.2.5, *logical story units* can hardly reach the scene boundaries and our attempts[2] to automatically merge consecutive LSUs if both socially and lexically close enough did not prove to be sufficiently effective. Manual annotation of scene boundaries turns out to be quite straightforward and does not require much time (about 10% of the film real time duration). The reference scene boundaries, as manually annotated, were thus used for estimating interactions from speakers co-occurrence.

Moreover, automatically detecting "who spoke when" in a movie is quite challenging: as emphasized in Section 3.5, speaker diarization turns out to be especially tricky when applied to TV series, often containing many speakers talking in adverse acoustic conditions (sound effects, background music...). Despite the benefits of the multi-modal approaches we detailed in Section 3.5, the error rates obtained when applying speaker diarization tools to TV series remain too high (about 50%) to serve as a reliable basis for building interaction networks. As we said, the speakers are thus manually indicated, by labeling the subtitles according to the corresponding speakers. This annotation step is much more time-consuming than for scene delimitation, requiring in average as much time as the real duration of each film.

Nevertheless, though much more relevant than larger-grained units, the scene used as a way of capturing the verbal interactions between characters may result in weak, sometimes irrelevant, interactions: if being at the same place at the same time is usually required to consider that several persons interact, it is rarely sufficient. Fig. 4.4 shows two consecutive dialogues extracted from the TV serial *House of Cards*, and belonging to the same scene. Three speakers are involved, but without any interaction between the second (*D. Blythe*) and the third (*C. Durant*) ones. The first speaker (*F. Underwood*) is talking to *D. Blythe* in the first sequence, then is moving to *C. Durant* and starts discussing with her.

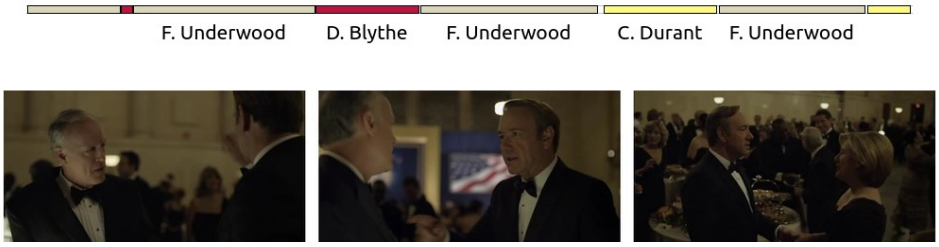


Figure 4.4 – *Two consecutive dialogue sequences within the same scene.*

The resulting interaction triangle, based on this scene co-occurrence, is shown on Fig. 4.5: *D. Blythe* and *C. Durant* are linked whereas they are not involved in any direct verbal interaction. One way of addressing this issue would be to consider even smaller units than scenes, but such a notion of a "sub-scene" may be confusing and difficult to define objectively. Another way of facing this problem would be, instead of glob-

[2]Not reported in this work.

ally considering the scene unit, to build the verbal interactions upon the sequence of utterances in each scene.

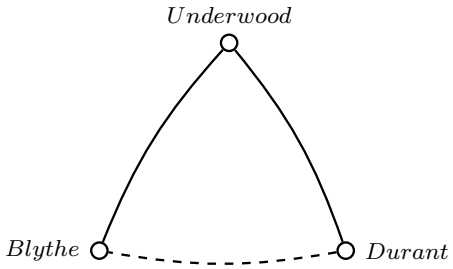


Figure 4.5 – *Co-occurrence network based on the sequence shown on Fig. 4.4. The interaction wrongly introduced is drawn in dash line.*

We now focus on relationships defined in a *strong sense*, as based on personal verbal interactions between characters. The resulting network can therefore be considered as a *conversational network*, in contrast to the co-occurrence network of characters described in (Weng et al., 2007, 2009) and used in (Ercolessi et al., 2012).

4.3.2 Sequential estimate of verbal interactions

Instead of globally considering the scene unit, we choose to tackle this problem by identifying the verbal interactions from the sequence of speech turns in each scene, once manually labeled according to the corresponding speakers. In order to estimate the verbal interactions from the single sequence of utterances, we apply four basic heuristics:

- **Rule *(1)* : Surrounded speech turn.** We consider that a speaker s_2 is talking to another speaker s_1 if he is speaking both after and before him, resulting in a speech turns sequence $s_1 s_2 s_1$, where each speech turn is labeled according to the corresponding speaker. Fig. 4.6 shows the subgraph resulting from the application of Rule *(1)* to the speech turns sequence shown on Fig. 4.4, where each edge is labeled according to the number of times each speaker is considered as talking to another one.

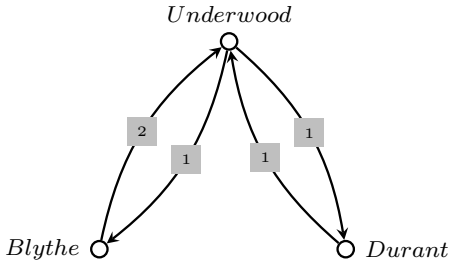


Figure 4.6 – *Number of directed links resulting from the application of Rule (1) to the speech turns sequence shown on Fig. 4.4.*

- **Rule *(2)* : Starting and ending speech turns.** This rule aims at processing the first and last utterances of each sequence $s_1 s_2 ... s_3 s_4$ of speech turns, by adding two links $s_1 \rightarrow s_2$ from the first to the second speaker and $s_4 \rightarrow s_3$ from the fourth to the third one. The network resulting from the application of Rule *(2)* to the sequence of Fig. 4.4 is shown on Fig. 4.7.

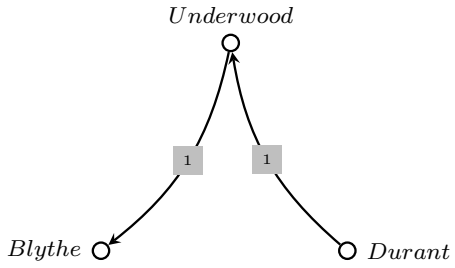


Figure 4.7 – *Number of directed links resulting from the application of Rule (2) to the speech turns sequence shown on Fig. 4.4.*

The last two rules are introduced to process ambiguous sequences of the type $s_1 s_2 s_3$, where three consecutive speech turns originate in three different speakers: in such cases, the second speaker might be stated as talking to the first one as well as to the third one, or even to both of them. However, such speech turn sequences can often be disambiguated by focusing on speakers involved both before and after the considered sequence.

- **Rule *(3)* : Local disambiguation.** We distinguish two variants of this rule. Rule *(3a)* applies when the second speaker appears before the sequence, but not after, as in $(s_2) s_1 s_2 s_3 (s_4)$. We then consider s_2 is speaking with s_1 rather than with s_3. Symmetrically, Rule *(3b)* concerns the case when the second speaker appears after, but not before the sequence, as in $(s_0) s_1 s_2 s_3 (s_2)$, and is therefore assumed to speak to s_3.

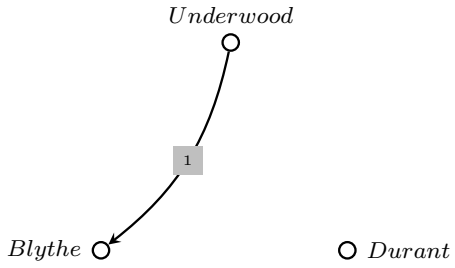


Figure 4.8 – *Number of directed links resulting from the application of Rule (4) to the speech turns sequence shown on Fig. 4.4.*

- **Rule *(4)* : Temporal proximity.** When the second speaker is involved in the conversation both before and after the ambiguous sequence, as in $(s_2) s_1 s_2 s_3 (s_2)$, we consider the ambiguous speech turn to be intended for the speaker whose utterance is temporally closer. In the sequence shown on Fig. 4.4, the fifth, ambiguous

utterance would then be hypothesized as intended for the first speaker *D. Blythe*, resulting in the additional link shown on Fig. 4.8. The same Rule *(4)* is applied when the speaker s_2 is not involved in the immediate conversational context.

Fig. 4.9 shows the total amount of directed interactions between any two speakers involved in the scene shown on Fig. 4.4, after applying Rules (1–4). On the left-hand part 4.9a of the figure, the weight of the links is the number of times one speaker is talking to another one; on the right-hand part 4.9b, the weight is computed as the total duration of the interaction.

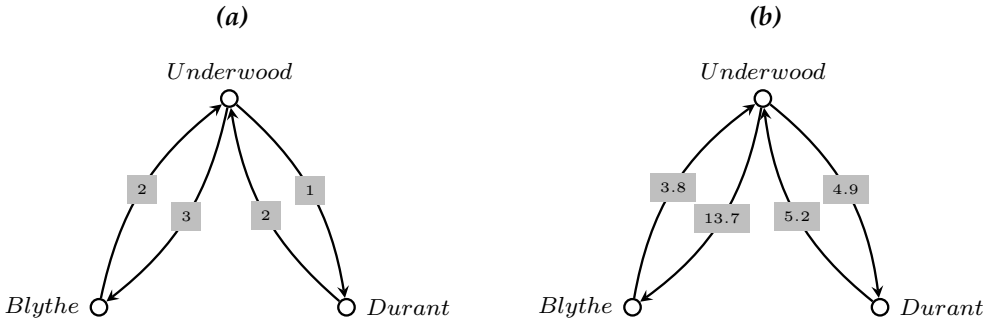


Figure 4.9 – *Directed links resulting from the application of Rules (1–4) to the speech turns sequence shown on Fig. 4.4, with weights corresponding either to the number of interactions (4.9a) or to interaction time in seconds (4.9b).*

We now describe the algorithm we use to build a dynamic network of interacting speakers able to capture the evolution of the narrative content over time.

4.4 Dynamic conversational network for plot modeling

4.4.1 Narrative smoothing

We obtain the total amount of interaction $h_{ij}^{(t)}$ between the speakers s_i and s_j in the t^{th} scene by summing up the amount of speech flowing in the scene t from s_i to s_j and from s_j to s_i, resulting in an undirected local interaction amount, possibly zero, expressed in seconds.

As stated in Section 4.1, we would like to get an instantaneous measurement of the strength of any relationship at any moment, but from the successive partial views of the underlying network that the narrative provides us. Intuitively, a particular relationship may be considered as especially important at some point of the story if the involved characters both speak frequently and a lot to each other: the time interval needed before the interaction is reactivated in the narrative is expected to be short, and the interaction time to be long whenever the relationship is active in the plot.

Four possible states have to be considered when monitoring a single relationship over time: (1) the relationship is active in the current scene; (2) it has been active in the

story and will be active again later; (3) it was active before, but will no longer be active in the narrative; and (4) it has not yet been active in the narrative.

(1) Relationship currently active in the story

The first case is the simplest one: each time the interaction occurs, its strength can be estimated in a standard way as the duration of the interaction, expressed in seconds: in any scene t where speakers s_i and s_j are hypothesized as talking to each other, the instantaneous weight of their relationship $w_{ij}^{(t)}$ is estimated as follows:

$$w_{ij}^{(t)} = h_{ij}^{(t)} \tag{4.1}$$

where $h_{ij}^{(t)}$ denotes the interaction time, expressed in seconds, between the i^{th} and j^{th} speakers in scene t.

The last three cases are much trickier. Between two consecutive occurrences of the same relationship in the story, it would be tempting to consider that the relationship is still (resp. already) active if it is recent (resp. imminent) enough at each moment considered.

According to the time-slice framework described in Section 4.2, as long as the relationship is present in the observation window of the network over time, it is stated as active, and inactive as soon as no longer observed. A smoother alternative based on temporal decay is used in (Mutton, 2004).

As emphasized in Section 4.2, such a way of handling the past and future occurrences of the relationship is inappropriate for most TV serials. Some interacting characters may be absent from the narrative for an undefined period of time but still linked in the underlying network, as confirmed by the fact that the last state of the relationship is generally used as a starting point when the characters are re-introduced in the story. Indeed, the temporalness of the narrative should affect a relationship only when at least one of the involved characters interacts with others after and/or before the relationship is active: the relationship between two characters should only get weaker if they interact separately with others before interacting again with one another.

In order to perform such a *narrative smoothing*, we introduce two auxiliary quantities to handle the scenes where the two characters do not interact. First, $\Delta_{ij}^{(l)}(t)$ is the *narrative persistence* in scene t of the last occurrence l of the relationship between speakers s_i and s_j:

$$\Delta_{ij}^{(l)}(t) = h_{ij}^{(l)} - \sum_{t'=l+1}^{t} \sum_{k \neq i,j} \left(h_{ik}^{(t')} + h_{jk}^{(t')} \right) \tag{4.2}$$

This measure $\Delta_{ij}^{(l)}(t)$ corresponds to the net balance between the duration $h_{ij}^{(l)}$ of the last interaction between the two characters s_i and s_j and the conversational time

(represented by the double sum) s_i and s_j have devoted separately to other characters s_k since then.

Symmetrically, $\Delta_{ij}^{(n)}(t)$ is the *narrative anticipation* in scene t on the next occurrence n of the relationship between speakers s_i and s_j:

$$\Delta_{ij}^{(n)}(t) = h_{ij}^{(n)} - \sum_{t'=t}^{n-1} \sum_{k \neq i,j} \left(h_{ik}^{(t')} + h_{jk}^{(t')} \right) \tag{4.3}$$

(2) Relationship between two narrative occurrences

We then define the instantaneous weight $w_{ij}^{(t)}$ of the relationship between the speakers s_i and s_j in any scene t occurring between two consecutive occurrences of their relationship as:

$$w_{ij}^{(t)} = \max \left\{ \Delta_{ij}^{(l)}(t), \Delta_{ij}^{(n)}(t) \right\} \tag{4.4}$$

If neither of the two characters speaks to others before they interact again with one another, $w_{ij}^{(t)} = \max \left\{ h_{ij}^{(l)}, h_{ij}^{(n)} \right\}$ and the last (resp. next) occurrence of the relation is considered as still (resp. already) fully present in the network, whatever the number of intermediate scenes the narrative introduces in-between to focus on other plot substories.

(3) Relationship after its last narrative occurrence

The weight of the relationship between the i^{th} and j^{th} speakers in any scene t occurring after its very last occurrence in the narrative is expressed as follows, provided that one of the two characters remains involved in the story by interacting with others:

$$w_{ij}^{(t)} = \Delta_{ij}^{(l)}(t)$$

(4) Relationship before its first narrative occurrence

Symmetrically, the weight of the relationship between the i^{th} and j^{th} speakers in any scene t occurring before its first occurrence in the story is computed as follows, as long as one the two characters has already been shown as interacting with other people:

$$w_{ij}^{(t)} = \Delta_{ij}^{(n)}(t)$$

In the very last case, when neither of the two characters is still (resp. already) active, the weight w_{ij} is set to $-\infty$.

We then normalize the weights of the interactions linking any couple of characters in some scene t. We use the following formula, resulting in an undirected graph $\mathcal{G}^{(t)}$,

capturing the instantaneous state of the social network that the story sequentially unveils:

$$n_{ij}^{(t)} = \frac{1}{1 + e^{-\lambda w_{ij}^{(t)}}} \tag{4.5}$$

where $n_{ij}^{(t)}$ is the normalized weight of the relationship between the speakers s_i and s_j. The choice of the sigmoid function for such a normalization purpose both allows to get weights ranging from 0 to 1 and to simulate the way the past and future states of a relationship in the narrative could influence its current state at some point t. The parameter λ is a parameter of sensitivity to the past and future states of the network and was set to $\lambda = 0.01$ (high values imply low dependence on the future and past states).

4.4.2 Narrative smoothing illustrated

Fig. 4.10 shows excerpts of four consecutive scenes in *House of Cards*, involving five individuals. The first two of them, namely *Francis Underwood* and his wife *Claire*, interact with each other in the first and last scenes (red border) respectively during 30 and 20 seconds, whereas Claire interacts in-between 40 seconds with another person in the second scene (green border) and two other people are talking to one another in the third scene during 50 seconds.

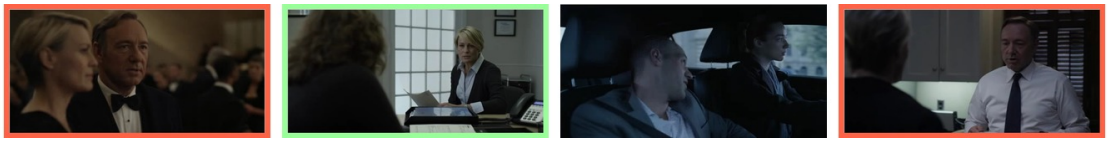


Figure 4.10 – *Example of application of the weighting scheme to a specific relationship.*

In the first and fourth scenes, Claire and Francis are interacting with each other: according to Equation 4.1, we then set the weights of their relationship to the corresponding interaction times, respectively 30 and 20 seconds.

In the second scene, the last interaction between Claire and Francis is on the one hand weakened by the separate interaction of Claire with someone else during 40 seconds: the resulting *narrative persistence* of the relationship between Francis and Claire then amounts to $\Delta_{12}^{(1)}(2) = 30 - 40 = -10$ (Equation 4.2).

On the other hand, the *narrative anticipation* on the next occurrence of the relationship between Francis and Claire then amounts to $\Delta_{12}^{(4)}(2) = 20 - 40 = -20$ (Equation 4.3), resulting in an instantaneous weight $w_{12}^{(2)} = \max\{-10, -20\} = -10$ in the second scene.

In the third scene, neither of the two characters is involved: the narrative persistence of their relationship is unchanged, but the narrative anticipation then increases to 20, because no interfering character separates at this point Francis and Claire from their next interaction in the fourth scene. We then have $w_{12}^{(3)} = \max\{-10, 20\} = 20$ and the full resulting sequence of unnormalized, instantaneous weights for the relationship between Claire and Francis is then (30, -10, 20, 20) at the four considered moments.

4.5 Experiments and results

In this section, we evaluate the whole framework we introduced for building a reliable and informative dynamic social network of interacting characters in TV serials. We first evaluate the basics heuristics we introduced in Subsection 4.3.2 to infer the interacting speakers from the sequence of speech turns, once manually annotated according to the corresponding speakers. We then qualitatively evaluate narrative smoothing, our graph extraction method, by comparing it to both types of methods described in Section 4.2. For this purpose, we focus on every TV serial of our corpus, and explore their plots from the dynamics of their underlying social network of characters. We both analyze the obtained networks from the perspective of the protagonists (nodes) and their relationships (links). We now describe the corpus subset we used when evaluating our methods.

4.5.1 Corpus subset

When designing our algorithms for estimating the interacting speakers, we used a more comprehensive subset of our corpus than when performing the tasks described in Chapter 3. For each of the three TV serial, the number of selected episodes is set so that the total film duration remains quite the same, whatever the TV serial. More specifically, the episodes are selected as follows:

- *Breaking Bad*: Season 1 (7 episodes) along with first 4 episodes of Season 2.
- *Game of Thrones*: Season 1 (10 episodes).
- *House of Cards*: first 11 episodes of Season 1.

Table 4.1 reports the main features of the resulting corpus: all the figures are computed from manual annotation.

As can be seen in Table 4.1, the spoken parts in each TV serial cover in average a bit less than half of the total duration of the films. Speech is more represented in *House of Cards* (coverage $\simeq$ 50% of the total time with 8,520 subtitles) than in the other two series.

Speech is uniformly distributed over the scenes, with in average more than 95% of the scenes containing at least one subtitle, which suggests that most social interactions are expressed verbally in these three TV serials.

Table 4.1 *– Corpus: main features.*

CORPUS	BB	GoT	HoC
# episodes	11	10	11
total duration (H:MM:SS)	8:25:04	8:28:51	8:21:42
speech coverage. (%)	39.5	43.8	50.7
# subtitles	6,182	6,998	8,520
# speaker occurrences	501	732	951
# scenes	206	249	390
% spoken scenes	94.7	97.2	97.2
# speakers/scene (avg.)	2.43	2.94	2.44
# speakers/scene (std. dev.)	1.22	1.52	1.14

Furthermore, the average number of speakers by scene remains quite low (ranging from 2.43 to 2.94 depending on the TV series), often resulting in simple patterns of verbal interactions properly handled by applying the basic heuristics described in Subsection 4.3.2.

We now turn to the evaluation of the accuracy of the four basic rules we use to estimate verbal interactions from the only sequence of speaker-labeled speech turns.

4.5.2 Conversational interactions

The evaluation of the way we estimate verbal interactions from the sequence of speech turns is performed in two ways: first, intrinsically, by measuring the performance of our method for achieving the task of estimating interactions; second, extrinsically, by measuring the reliability of the cumulative network in which the application of the four basic heuristics introduced in Subsection 4.3.2 results. We first describe the episode sample we annotated for this part of the evaluation process.

Episodes test sample

In order to evaluate the reliability of the methods introduced for estimating verbal interactions, a subset of test episodes for each of the three TV serials is selected. For each series, the subset of episodes considered is defined so that the distribution of the number of speakers per scene remains representative of the same distribution in the whole set of episodes.

Fig. 4.11 shows the distribution of the number of speakers by scene in the corpus subset we introduced in Subsection 4.5.1, containing about 10 episodes for each of the three TV serials. The distribution of the number of speakers by scene is shown both for each of the three series (dashed lines) as well as in each of the three episode samples we chose as test subsets (points) for each TV serial. The frequency of a scene involving n speakers is computed as the proportion of utterances belonging to scenes with n speakers, rather than the proportion of scenes with n speakers itself. The length of a scene

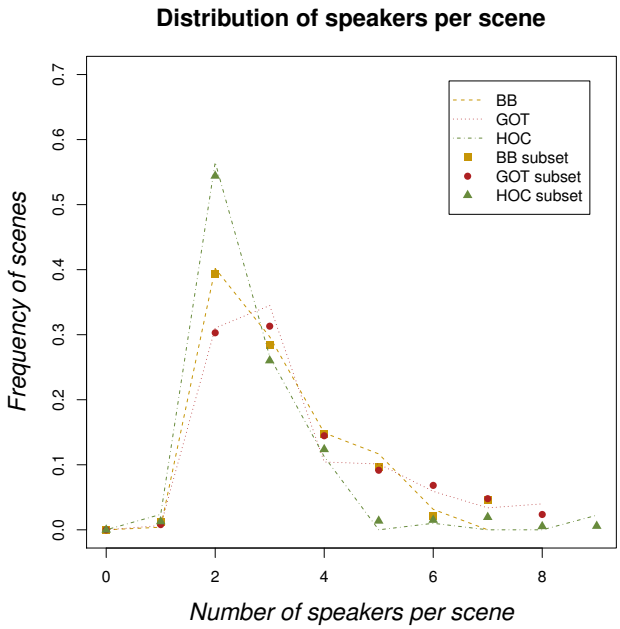


Figure 4.11 – *Distribution of speakers per scene: corpus (dashed lines), sample (points).*

increasing in average as the number of speakers grows, these two ways of computing the scene frequency are not exactly equivalent. A subset of episodes could for instance contain in proportion as many scenes with three speakers as in the whole corpus, but significantly shorter, artificially resulting in fewer complex patterns of speech turns and in better performances when estimating speaker interactions.

For each of the three resulting TV serial subsets, the same features as those computed in Table 4.1 are reported in Table 4.2.

Table 4.2 – *Test corpus: main features.*

TEST	BB subset	GOT subset	HOC subset
# episodes	4	3	3
episodes	4, 6, 10, 11	3, 7, 8	1, 7, 11
total duration (H:MM:SS)	2:59:31	2:37:17	2:29:23
speech coverage (%)	45.16	48.2	44.2
# subtitles	2,254	2,282	2,194
# speaker occurrences	202	233	231
# scenes	82	81	99
% spoken scenes	95.1	97.5	97.0
# speakers/scene (avg.)	2.46	2.88	2.33
# speakers/scene (std. dev.)	1.14	1.49	1.11

For each episode of the three test sets, each utterance is manually labeled according

to the speakers it is intended for. For monologues, where no specific listener is targeted, a special *null* label is introduced, and in case of multiple addressees for a single utterance, multiple labels were assigned.

Evaluation metrics

In estimating verbal interaction, the decision is made at the utterance level, by assigning to each utterance the speaker(s) it is intended for. The task then consists in categorizing every utterance among the available speaker classes, with multiple classes allowed if the utterance is labeled as addressed to multiple characters. Standard performance measures used in *information retrieval* in order to evaluate the multi-label categorization task can therefore be used as intrinsic evaluation metrics. More specifically, we perform intrinsic evaluation of the basic rules we introduced for estimating verbal interactions by using the following evaluation procedures discussed in (Tsoumakas et Katakis, 2006) for multi-label categorization:

- **Recall**:

$$R(\gamma, \mathbb{X}) = \frac{1}{|\mathbb{X}|} \sum_{y \in \mathbb{X}} \frac{|\gamma(y) \cap M(y)|}{|M(y)|}$$

  where $\mathbb{X}$ denotes the utterance set; $\gamma(y)$ denotes the set of interlocutor(s), possibly multiple, hypothesized for the utterance y, and $M(y)$ the set of reference interlocutors, as manually labeled, for the utterance y. In this context, recall is the average proportion, for every utterance y, of retrieved interlocutors among the reference ones.

- **Precision**:

$$P(\gamma, \mathbb{X}) = \frac{1}{|\mathbb{X}|} \sum_{y \in \mathbb{X}} \frac{|\gamma(y) \cap M(y)|}{|\gamma(y)|}$$

  In this context, precision corresponds to the average proportion, for every utterance y, of relevant interlocutors among the retrieved ones.

- **F-score**:

$$F(\gamma, \mathbb{X}) = \frac{2P(\gamma, \mathbb{X})R(\gamma, \mathbb{X})}{P(\gamma, \mathbb{X}) + R(\gamma, \mathbb{X})}$$

Besides intrinsic evaluation of the task of estimating verbal interactions, we also perform extrinsic evaluation of the resulting social network. Following the method described in (Agarwal et al., 2013) for a similar extrinsic evaluation purpose, we first convert into vectors by simple column concatenation the two adjacency matrices of the cumulative conversational networks resulting from the hypothesized and from the actual interactions. We then measure the similarity between the two resulting vectors. When interactions are weighted, the hypothesized and reference social networks are

compared by computing the euclidean distance and the cosine similarity between the two vectors of edge weights. When focusing on the mere fact that two characters verbally interact with each other whatever the interaction amount, we do not weight interactions, and we evaluate their similarity by computing the Jaccard index of the two sets of edges, both as hypothesized and as manually labeled. The measures of similarity between the hypothesized and reference networks are computed both when discarding the first and last utterances of each scene and when considering every speech segment: the first utterance of the next scene is sometimes slightly anticipated at the very end of the current one, possibly resulting in irrelevant interactions.

By using both intrinsic and extrinsic evaluation metrics, it is possible to measure the performance of the rules we used for sequentially estimating the verbal interactions.

Evaluation results

In evaluating the performance of the rules used for sequential estimate of verbal interactions, we follow a step-by-step process, by successively using in conjunction to the first, most robust rule, the third remaining ones. Fig. 4.12 shows the change, both in coverage and performance (intrinsic F-score and extrinsic similarity measures), when applying a more and more comprehensive set of rules, from the single first one, denoted *(1)*, to the whole four rules, denoted *(1–4)*.

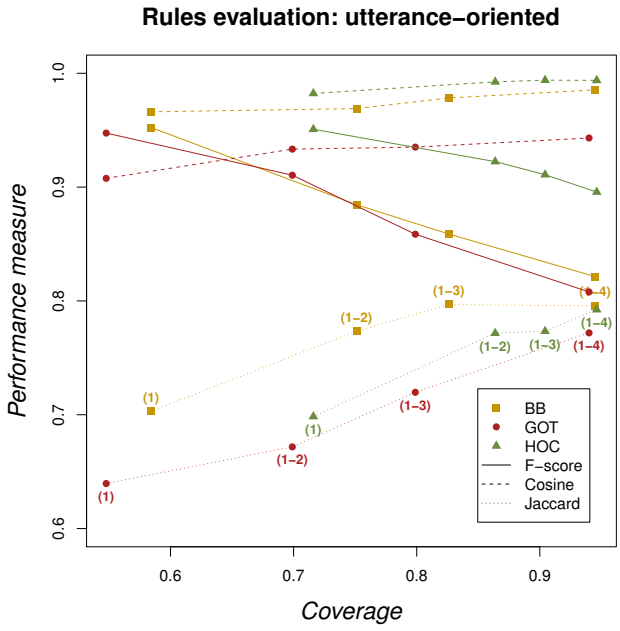


Figure 4.12 – *Step-by-step evaluation or the rules used for sequentially estimating verbal interactions.*

Not surprisingly, the more rules are used, the more interactions are hypothesized. As reported in Table 4.3, the very basic first rule (*surrounded speech turn*) allows in average to hypothesize interlocutors for 62% of the spoken segments. When the whole set

of rules is used, decisions are made for 94% of the utterances. The remaining utterances correspond to soliloquies or isolated utterances.

Table 4.3 *– Evaluation of the joint use of the rules applied for sequentially estimating speakers interactions. The cosine, L2, Jaccard measures are both computed when discarding the first and last segments of each scene, or not.*

Rules	Evaluation metrics	TV serial			
		BB	GOT	HOC	**avg.**
(1)	**coverage**	0.58	0.55	0.72	0.62
	F-score	0.95	0.95	0.95	0.95
	Precision	0.96	0.95	0.95	0.95
	Recall	0.94	0.94	0.95	0.94
	Jaccard sim.	0.70	0.64	0.70	0.68
	cos. sim.	0.97 - 0.96	0.91 - 0.89	0.99 - 0.99	0.96 - 0.95
	L2 dist.	0.26 - 0.29	0.43 - 0.47	0.19 - 0.16	0.29 - 0.31
(1–4)	**coverage**	0.94	0.94	0.95	0.94
	F-score	0.82	0.81	0.90	0.84
	Precision	0.84	0.82	0.90	0.85
	Recall	0.80	0.80	0.89	0.83
	Jaccard sim.	0.80	0.77	0.79	0.79
	cos. sim.	0.99 - 0.98	0.94 - 0.93	0.99 - 0.99	0.97 - 0.97
	L2 dist.	0.17 - 0.20	0.34 - 0.37	0.11 - 0.13	0.21 - 0.23

More surprisingly, as can be seen both on Fig. 4.12 and in Table 4.3, the additional rules *(2–4)* introduce more and more mistakes when hypothesizing interlocutors at the utterance level, resulting in a lower and lower F-score, but in the meantime, the extrinsic evaluation measures of the resulting network are improving: as can be seen in Table 4.3, while the F-score decreases in average from 0.95 to 0.84, the euclidean distance between the hypothesized and the reference networks decreases from 0.29 to 0.21 if the first and final utterances of every scene are discarded, or from 0.31 to 0.23 if not.

Such an inconsistency suggests that errors made locally when assigning each utterance to the addressed characters do not deteriorate the reliability of the resulting conversational network: indeed, only a small proportion of the errors made at such a local level (utterance-level F-score amounting to zero) introduces irrelevant links in the resulting global network (14.08% for *Breaking Bad*, 13.43% for *Game of Thrones* and 5.34% for *House of Cards*). Moreover, some errors made at the utterance-level by using more and more covering rules allow to retrieve interactions that would otherwise have been missed, or improperly weighted, by carefully applying the only rule (1): the additional rules *(2–4)* tend to introduce correct interactions, but at wrong places, and finally result in more reliable conversational networks, with more actual relationships captured and more representative link strengths when weighting interactions. Though basic, the four heuristics introduced in Subsection 4.3.2 turn out to be very effective when building cumulative conversational network. The reliability of these heuristics is qualitatively illustrated in Appendix B by showing the resulting cumulative networks,

as built upon the first two seasons of each of our TV serials.

We now qualitatively evaluate narrative smoothing, the method we use to build the dynamic conversational network from the interactions between speakers.

4.5.3 Narrative smoothing

The protagonists

We first base our analysis on the protagonists of the considered TV serials, *i. e.* the nodes in the corresponding extracted social networks. We present only a small number of results, which concern characters of particular interest. We characterize them using the *node outgoing strength*, a generalization of the node degree defined as the sum of the weights of the links originating from the considered node. In our case, weights are based on spoken interaction durations, so the strength of a character is related to how much and how frequently he speaks to others.

We first focus on Walter White, the main character of *Breaking Bad*, and Tuco Salamanca, one of the drug dealers with whom he is in business. When considering the cumulative network of *Breaking Bad* as illustrated on Fig. B.1, *i. e.* the temporal integration over the first 20 episodes, the strength of Walter White (his total interaction time with others) is about twenty times as large as the strength of Tuco: 12,332 seconds for Walter (rank 1) *vs.* 590 for Tuco (rank 11)[3].

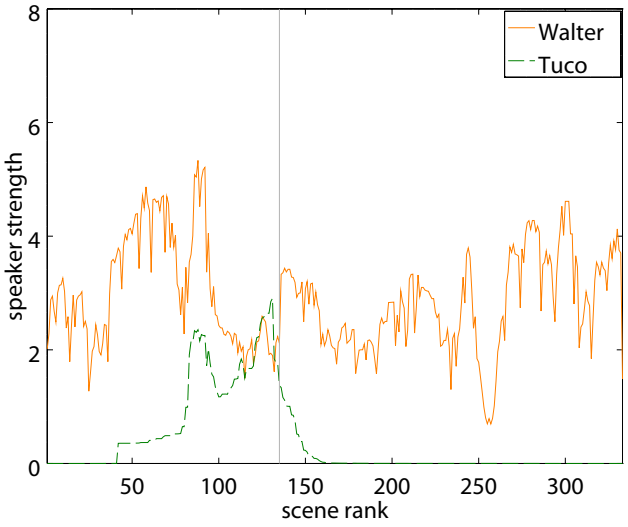


Figure 4.13 – *Strength of two important characters in* Breaking Bad, *plotted as a function of the scenes.*

By comparison, Fig. 4.13 displays the evolution of their instantaneous strength, obtained with our narrative smoothing method, as a function of the scenes ordered

[3]Node size on Fig. B.1 is not linearly proportional to strength.

chronologically. This leads us to a completely different vision of Tuco's role in the plot. As Fig. 4.13 shows, from scene 100, his importance tends to increase and even overcomes the importance of the main protagonist for some time, before suddenly decreasing after scene 130. This clearly corresponds to a subplot, or a short narrative episode, ending with Tuco's death, at the end of scene 135 (vertical line on Fig. 4.13).

We now switch to Daenerys Targaryen and Tyrion Lannister, two major protagonists of *Game of Thrones*. Fig. 4.14 shows how their strength evolves over the first two seasons of the series, again as a function of the chronologically ordered scenes, and illustrates the limitations of time-windowing approaches.

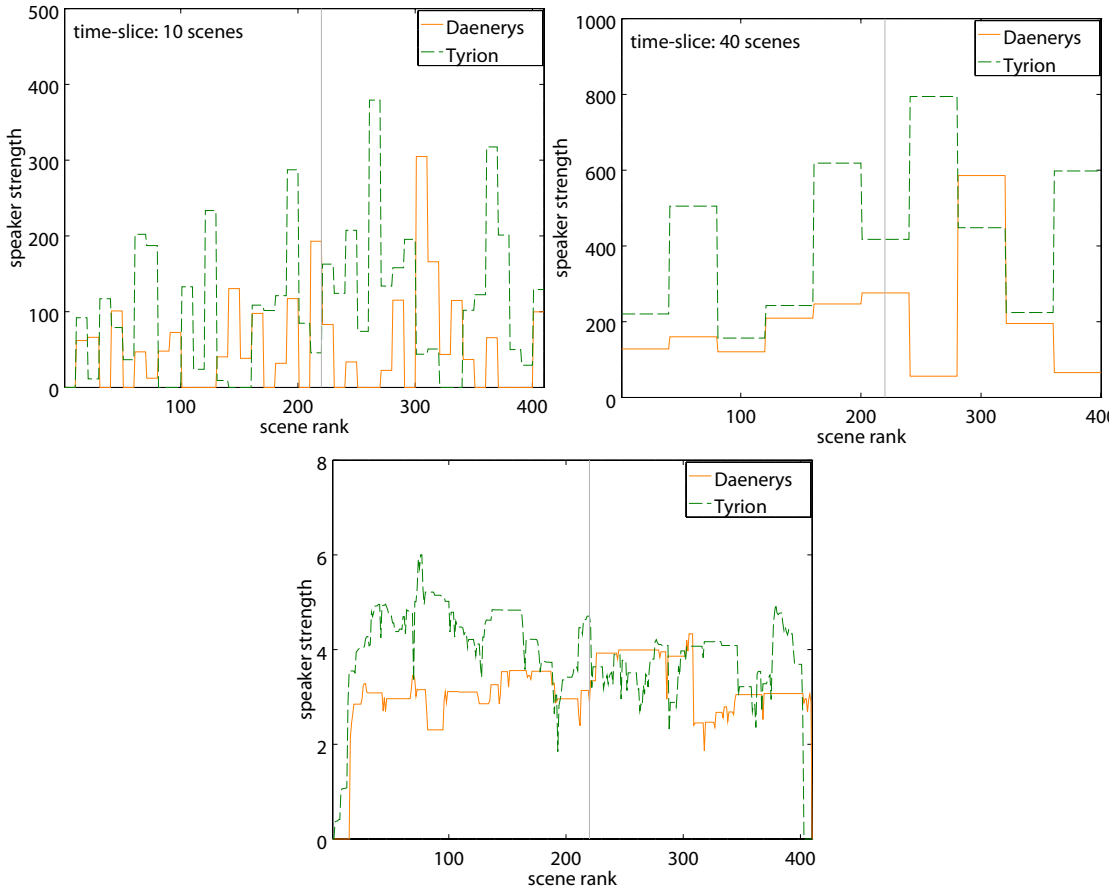


Figure 4.14 – *Strength of two major characters of* Game of Thrones *plotted as a function of the chronologically ordered scenes. Top:* 10 *and* 40 *scenes time-slices. Bottom: narrative smoothing.*

The first two plots, on the top of Fig. 4.14, were obtained through the use of fixed-size observation windows, set to 10 scenes (around half an episode) for the first and 40 (about two episodes) for the second. The last plot, on the the bottom, relies on our narrative smoothing method. The appearance of Daenerys' storyline onscreen has a relatively slow pace in these seasons (Fig. 4.2) and as can be seen, when the window is too narrow, this creates noisy, irrelevant measurements of her narrative importance

(top left on Fig. 4.14). It appears very unstable because her storyline alternates with many others on the screen. A wider observation window (top right on the same figure) is more likely to cover successive occurrences of Daenerys in the narrative, but, unlike our narrative smoothing method, prevents us from locating precisely the scenes responsible for Tyrion's current importance: for instance, a local maximum in Tyrion's strength is reached at scene 220 (third plot on Fig. 4.14), just after a major narrative event took place – the nomination of Tyrion as the King's Counselor (vertical line). Such an event remains unnoticed when accumulating the interactions during too large time-slices (second plot on Fig. 4.14), but is well captured by our approach.

Fig. 4.14 also reveals an important property of our way of building the dynamic network. Because the past (resp. future) occurrences of a particular relationship are still (resp. already) active as long as the involved characters do no interact with others in the meantime, the respective strengths of the main characters of the story appear remarkably balanced. Whereas Tyrion looks much more central than Daenerys in the time-slice based dynamic networks, whatever the size of the observation window, Daenerys is nearly as central as Tyrion in the network based on our narrative smoothing method: few of her acquaintances are shown onscreen as interacting with others. On the opposite, the story focuses more frequently on Tyrion, but also on separate interactions of his usual interlocutors, weakening his instantaneous strength (especially after scene 220): the dynamic strength, as computed after applying narrative smoothing, does not reduce to a global centrality measure, but also corresponds to a more local property, that measures how exclusively a character is related to his/her social neighborhood.

Our results confirm that cumulative networks, by neglecting the temporal dimension, tend to completely miss punctual changes in the importance of certain characters relatively to the plot. The time-slice based methods can handle the network dynamics, however our observations illustrate that they cannot properly tackle the narrative issue we described in Subsection 4.2.2. The choice of an appropriate time window is a particularly sensitive point. By comparison, narrative smoothing captures the state of a relationship at any moment of the plot, using a time scale which directly depends on the narrative pace of the considered series. This allows to finely evaluate the degree of instantaneous involvement of any character in the plot.

The relationships

We now consider relationships between pairs of characters, instead of single individuals. We characterize each relation depending on its weight, *i. e.* the amount of time the characters talked to each other, either cumulated over time-slices, possibly consisting of the whole set of episodes, or smoothed with respect to the narrative. Like for the protagonists, we focus on relationships of particular interest.

Let us consider two relationships in *House of Cards*, representative of two substories: the first one corresponds to a narrative sequence in the storyline related to the main character Francis Underwood – his fight with a former ally, the unionist Martin Spinella; the second one is a similar subplot, but related to a secondary character, not as frequently present in the narrative, the journalist Lucas Goodwin, who requests the help of the hacker Gavin Orsay to investigate on Francis. Though locally important in

these two substories, neither of these relationships lasts long enough to be noticed in the cumulative network, as resulting from the first two seasons of the series (shown on Fig. B.3): the interaction time amounts to 562 seconds for the relation between Francis and Martin, and to 294 seconds for the relation between Gavin and Lucas. These total interaction times remain quite small compared to the central relation between Claire and Francis, amounting to 2,319 seconds.

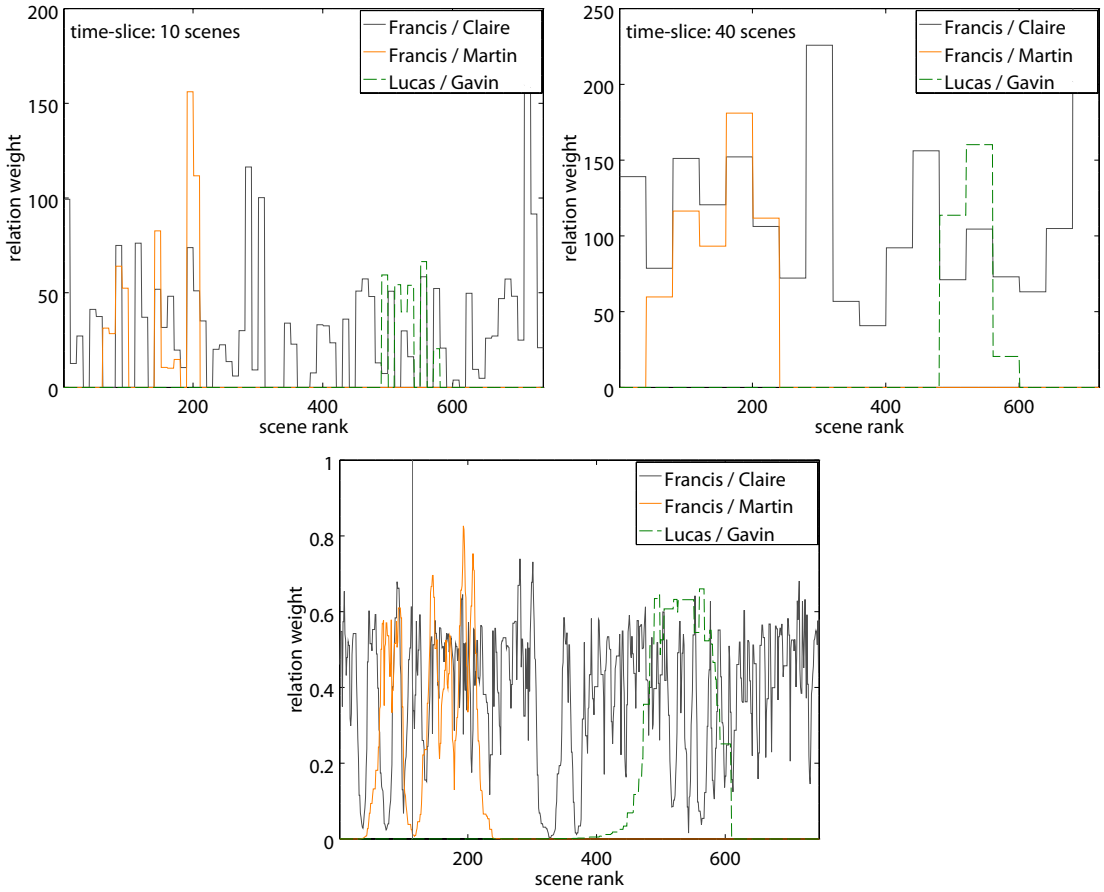


Figure 4.15 – *Weight of three relationships between five characters of House of Cards plotted as a function of the chronologically ordered scenes. Top:* 10 *and* 40 *scenes time-slices. Bottom: narrative smoothing.*

Nonetheless, once plotted as a function of the chronologically ordered scenes (Fig. 4.15), the respective weights of these relationships in the narrative look quite different, whatever the weighting scheme. Both substories, the one based on the relation between Francis and Martin and the one based to the relation between Lucas and Gavin, turn out to be locally as important as the long-term substory based on the relation between the two main characters Claire and Francis. However, all three ways of monitoring these relationships over time are not equivalent: agglomerating the interactions within short time-slices (top left plot on Fig. 4.15) makes us miss the continuity of Lucas/Gavin's substory, which occurs *in the narrative* at a slower rate than the substories related to

Francis. Conversely, large time-slices (top right plot on Fig. 4.15) allow to capture this substory, but agglomerate the two main stages of the relation Francis/Martin: before becoming an enemy, Martin is first an ally of Francis; these two parts in the relation correspond to well-separated stages in the narrative, that too large time-slices tend to merge. In contrast, such a breakpoint (materialized by a vertical line on the bottom plot of Fig. 4.15) is correctly captured when monitoring the relationship with narrative smoothing.

Let us go back once again to *Game of Thrones* and its complex plot. Fig. 4.16 focuses on two relationships between three characters: Catelyn Stark and Ned Stark on the one hand, Catelyn Stark and Tyrion Lannister on the other.

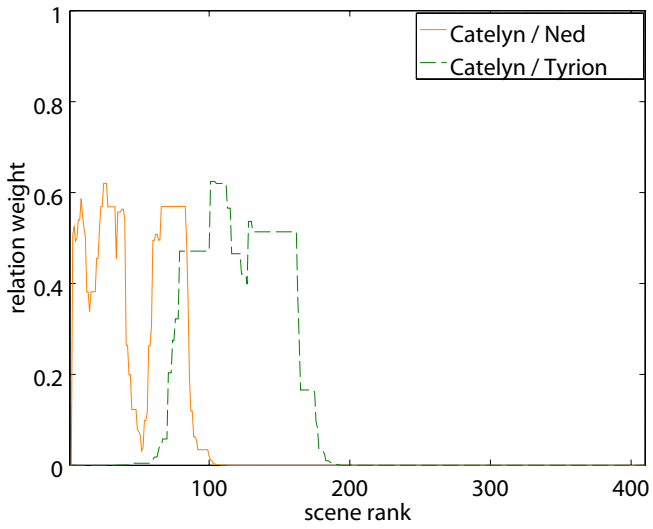


Figure 4.16 – *Weight of two relationships between three characters of Game of Thrones plotted as a function of the chronologically ordered scenes and based on narrative smoothing.*

Neither of these relationships would be considered as a major one from the cumulative graph at the end of the first two seasons (*cf.* Fig. B.2). Nevertheless, once dynamically considered, they both correspond to two successive substories in the first season of *Game of Thrones*. As can be seen, our narrative smoothing approach even allows to separate two steps in the relationship between Catelyn and Ned: the first step of their relationships takes place in Winterfell; Ned then leaves Catelyn there and goes on his own to King's Landing, freshly named as the King's Counselor (around scene 65), before Catelyn joins him there. Catelyn and Tyrion start interacting with each other after Catelyn leaves Kings Landing to Winterfell; their relationship is well preserved once monitored according to our method, though shown in the narrative at a quite slow and irregular pace.

Our results confirm that cumulative networks are inappropriate for capturing punctual substories supported by specific relationships. Moreover, though much more appropriate for such a task, time-slice approaches suffer from a major drawback: once fixed, the time slice cannot adapt to the variable rates at which the substories appear

in the narrative. By overcoming the narrative contingencies, our narrative smoothing approach allows to monitor more accurately over time any relationship, whatever the way the narrative focuses on it.

4.6 Conclusion

In this chapter, we made a new step towards TV serial summaries by modeling the plot content from the evolving social network of interacting characters. For this purpose, we described a novel way of monitoring over time the state of the relationships between characters involved in the usually complex plots of modern TV series.

The two methods previously used are the cumulative approach, consisting in integrating every relation over the whole considered period of time, and the time-slice approach, consisting in breaking down the timeline into smaller discrete chunks. The first one turns out to be relatively inefficient for investigating complex storylines and a dynamic perspective is more appropriate. The second one complies with this constraint, but defining an appropriate size for the observation window is a very difficult task and constitutes a major drawback: the plots of TV serials usually consist of parallel storylines shown sequentially onscreen at an unpredictable frequency. As a main consequence, the narrative disappearance in the current scene of some past relationship can usually not be interpreted as a real disappearance, which invalidates the time-slice approach.

To address this issue, we chose to smooth the narrative sequentiality, by considering that the relation between some speakers remains active as long as neither of them interacts with others; if so, such separate interactions result in a progressive dissolution of the past link. Symmetrically, the imminence of the next occurrence of the relationship has to strengthen the link.

We then evaluated on our corpus the rules we use for estimating the interacting speakers from the sequence of speaker-labeled speech turns: though possibly misleading punctually, they result in quite reliable estimates of the characters' relationships. We finally experimentally compared our way of building the dynamic network of interacting speakers, which we call *narrative smoothing*, to both mentioned approaches on the three TV serials of our corpus. Though exploratory and qualitative, our results show that our method leads to more relevant results than both other methods, when it comes to instantaneously monitoring the importance of a particular character or of a specific relationship at some point of the story.

As detailed in the next chapter, the way some characters temporarily aggregate at some point of the story in a community-like structure suggests that some narrative sequences result in the stabilization, possibly temporary, of certain areas in the network[4].

[4]We made publicly available on https://dx.doi.org/10.6084/m9.figshare.2199646 video files containing short animations of the dynamic networks in the three TV serials, along with every network file we used for our experiments.

By automatically detecting such a narrative stabilization of some groups of relationships, it is possible to split the story, either considered globally or from a specific character's perspective, into substories, without assuming a static, predefined, community structure.

Chapter 5

Character-oriented automatic summaries

Contents

5.1 Introduction

In this chapter, we detail the algorithms we use to automatically generate summaries of TV serials, along with the promising results we obtained when we performed a large-scale user study in a real-case scenario.

As detailed in Chapter 1, the most popular TV series nowadays are TV serials, defined by continuous, possibly complex plots, and usually spanning several seasons. Nonetheless, the narrative continuity of TV serials directly conflicts with the typical viewing conditions: because of modern viewing technologies, new TV serial seasons are commonly watched over short periods of time. As a major consequence, viewers are likely to be disengaged from TV serial plots, both cognitively and emotionally, when about to watch new seasons. Summaries of the past seasons of TV serials are then expected not only to effectively remind the plot content, but to emotionally re-engage viewers in the narrative.

For both purposes, summaries of TV serials can benefit from the empathetic relationship viewers are likely to have with some specific characters. Consequently, we designed an interactive summarization framework, where users are free to automatically generate summaries of any character's storyline, at any granularity level and of any duration they want. Such character-oriented summaries are first expected to fill each user's information needs: for instance, a specific user may only need summaries of secondary characters' storylines. Furthermore, by allowing users to focus in the first place on their favorite character(s), our summarization framework is expected to act as an effective trailer and make them want to know what comes next in the narrative.

In order to assess our summarization framework in the real-case scenario described above and detailed in Chapter 1, we specifically focused during the evaluation process on *Game of Thrones*, a few weeks before the new, sixth season of this popular TV serial was released.

Our main contributions in this chapter are the following:

- The first consists in making use of the dynamic social network of interacting characters, as built upon the narrative according to the approach detailed in Chapter 4, for capturing the specific dynamics of each character's storyline.
- The second consists in estimating the relevance of movie sequences in the context of summarization, by relying to some extent on the filmmaking stylistic patterns we introduced in Chapter 3.
- The third is the use of an additional criterion when applying the standard Maximal Margin Relevance algorithm (Carbonell et Goldstein, 1998) for building the final summary.
- Finally, the fourth is the user study we conducted, both to assess our method and to get valuable feedback for future work.

The rest of the chapter is organized as follows: in Section 5.2, we review the main

related works. In Section 5.3, we detail how the dynamic social network of interacting speakers detailed in Chapter 4 provides us with a way of modeling the dynamics of a specific character's storyline. In Section 5.4 we detail the way we estimate the relevance of each candidate unit, both from style and content-based features, along with the summarization algorithm. In Section 5.5, we describe the user study we performed and the main results we obtained.

5.2 Previous works

TV series content-based summarization. There is a limited amount of work that takes narrative content into account when creating TV series summaries. The most related works are probably (Tsoneva et al., 2007) and (Sang et Xu, 2010). As we detailed in Chapter 2, (Tsoneva et al., 2007) make use of both lexical and character-related features for automatically generating 10/15-minute video summaries of standalone TV series episodes. (Sang et Xu, 2010) make a further step towards using SNA for automatically generating character-based summaries of full-length movies and standalone episodes of TV series. Nonetheless, in both works, TV series episodes are considered independently as narratively self-consistent, and no attempt is made to summarize sequences of episodes, even in case of continuous plots.

Plot modeling in movies and TV series. As detailed in Chapter 4, (Weng et al., 2009) make use of SNA to automatically analyze the plot of full-length movies and TV series: the social network resulting from the agglomeration of every interaction between the characters is split into communities, before narrative breakpoints are hypothesized if the characters involved in successive scenes are socially distant. In (Ercolessi et al., 2012), a similar network of interacting speakers is used, among other features, for clustering into storylines the scenes of standalone episodes of two TV series. However, as in the summarization frameworks mentioned above, TV series episodes are considered as self-sufficient, and the analysis assumes a static structure of interacting characters. Similarly, (Tapaswi et al., 2014) focus on the interactions between the characters of TV series to build a visual, concise representation of the plot, but still for single episodes. In (Friedland et al., 2009a), the speech turns, among other features, are used to design a navigation tool for browsing sitcom episodes.

Stylistic patterns in movies. As reported in Chapter 4, (Guha et al., 2015) adopt a stylistic perspective for the plot modeling purpose, by automatically detecting the typical three-act narrative structure of Hollywood full-length movies. Most of the saliency-oriented techniques detailed in Chapter 2 can also be used to automatically analyze and even summarize fictional films, including TV series episodes. The video summarization scheme introduced in (Ma et al., 2002) is based on a content-independent attention model: some of the features used are closely related to filmmaking techniques commonly used to make viewers focus on specific sequences, such as shot size and music. (Hanjalic et Xu, 2005) investigate low-level features, some of them based on film grammar, like shot frequency, for modeling the emotional impact of videos, and especially full-length movies. Such low-level features, related to stylistic patterns used

in filmmaking, are widely used in automatic trailer generation. For instance, (Smeaton et al., 2006) make use, among other low-level features, of shot length and camera movement to isolate action scenes for later insertion in action movie trailers. Similarly, (Chen et al., 2004) introduce a way of automatically generating trailers and previews of action movies by relying on shot tempo.

However, none of these works, either content or saliency-oriented, differentiates between full-length movies and TV series episodes, which are always considered as self-sufficient from a narrative point of view. At such an episode scale, stylistic patterns are probably effective enough to reliably isolate salient sequences, and plot modeling can rely on reasonable hypotheses, as in (Weng et al., 2009) and (Ercolessi et al., 2012), about the distribution of the characters into static communities defined once and for all. Significantly, the most represented TV series genre in these works is the classical one, with standalone episodes and recurring characters.

In contrast, our focus is on TV serials, by far the most popular genre nowadays, defined by their continuous plots. At this much larger scale of dozens of episodes considered globally as developing a single plot, possibly split into multiple, parallel storylines, we cannot rely on the only stylistic patterns to isolate relevant sequences; furthermore, plot modeling requires a dynamic perspective and excludes every assumption about a stable and static community structure that the narrative would only unveil and never impact.

5.3 Modeling characters' storylines

In this section, we describe the way we model each character's storyline from the dynamic social network of interacting characters we introduced in Chapter 4. We then describe the weighting scheme we use to estimate the social relevance of a specific video sequence.

5.3.1 Narrative episode

In order to estimate the relevance of each candidate sequence, we first automatically segment the storyline associated to a specific character into *narrative episodes*. In any narrative, the story of a specific character usually develops sequentially and advances in stages: each narrative episode is defined as a homogeneous sequence where some event directly impacts a specific group of characters located in the same place at the same time. Though such a notion of narrative episode may be defined at different levels of granularity, such sequences are often larger than the formal divisions of books in chapters and of TV serials in episodes. Examples of such character-based narrative episodes in *Game of Thrones* could be: "Theon Greyjoy rules Winterfell"; "Arya Stark captive in Harrenhal"; "Jaime Lannister's journey in Dorne to rescue Myrcella"... The segmentation of each character's storyline into narrative episodes aims at building a summary able to capture the dynamics of the plot and is performed as follows.

We first build the weighted, undirected dynamic social network of interacting characters over the whole TV serial according to the method introduced in Chapter 4. As detailed in Section 4.4, the dynamic network is built upon the speech turns and scene boundaries, once manually annotated, and is based on a smoothing method that provides us with an instantaneous view of the state of any relationship at any point of the story, whether the related characters are interacting or not at that moment. As a result, the full, smoothed, social neighborhood of a specific character is always available in any scene t.

Based on this dynamic network $\mathcal{G}^{(t)}$, we define $\mathbf{r_t}$ as the relationship vector of a specific character in scene t: $\mathbf{r_t}$ contains the weights in scene t of all of his/her relationships and captures the instantaneous state of the neighborhood of the corresponding node. By construction, such weights range between 0 and 1 once normalized after applying Equation 4.5: high weights correspond to those of the character's relationships that are both long and frequent around the scene considered. For example, the following relationship vector $\mathbf{r_{63}}$, extracted from the narrative episode denoted above as "Jaime Lannister's journey in Dorne to rescue Myrcella", contains the weights of the most important relationships of the *Game of Thrones*' character Jaime Lannister, as computed in the 63$^{\text{rd}}$ scene where he is verbally involved:

$$\mathbf{r_{63}} = \begin{pmatrix} \text{Doran Martell} & [0.55] \\ \text{Myrcella Baratheon} & [0.53] \\ \text{Trystane Martell} & [0.51] \\ \text{Bronn} & [0.51] \\ & \vdots \end{pmatrix} \tag{5.1}$$

where the components are re-arranged in decreasing order.

For each character, we then compute the distance matrix $\mathcal{D}$, where $d_{t,t'}$ is the normalized euclidean distance between the character's relationship vectors $\mathbf{r_t}$ and $\mathbf{r_{t'}}$ in scenes t and t'. Because each narrative episode is defined as impacting a limited and well-identified group of interacting characters, the relationships of a character are expected to stabilize during each narrative episode, and to change when a new one occurs. Fig. 5.1 shows the matrix $\mathcal{D}$ for two *Game of Thrones*' major characters, Jaime Lannister (Subfig. 5.1a) and Arya Stark (Subfig. 5.1b). The matrices respectively show for these two characters the distance between their successive neighborhoods over the first five seasons of *Game of Thrones*. In this matrix, the time steps are the scenes, chronologically ordered, and for the sake of clarity, we build the matrix only upon the scenes where the considered character is involved, even though the smoothing method we use provides a way of estimating the social neighborhood of the character in any scene, whether the considered character is present or not.

As can be seen from Fig. 5.1, the character's social environment alternates between stable periods and new configurations, as the story develops. For instance, as shown on Subfig 5.1a, the social environment of Jaime Lannister after scene 60 remains pretty much the same and looks quite different from his previous relationships. Indeed, the

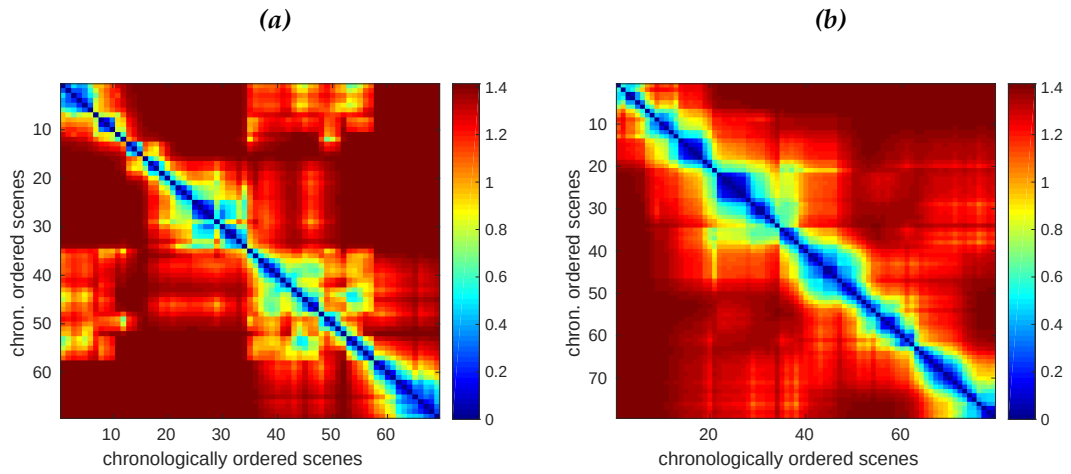


Figure 5.1 – *Matrix of distances between Jaime Lannister's (5.1a) and Arya Stark's (5.1b) social neighborhoods in any pair of scenes* (t, t') *in the first five seasons of* Game of Thrones.

social consistency of these scenes corresponds for Jaime to the narrative episode we denoted above as "Jaime Lannister's journey in Dorne to rescue Myrcella" (season 5). Similarly, as shown on Subfig 5.1b, between scenes 38 and 51, the social environment of Arya remains quite the same, suggesting that her storyline stabilizes in some narrative episode. Interestingly, other narrative episodes can also be observed in the matrix at larger (scenes 6–48) or smaller (scenes 21–26) scales, confirming the relative and multi-scale nature of the notion of narrative episode.

5.3.2 Optimal partitioning

Once the social distance matrix $\mathcal{D}$ is built over time for every character, we optimally partition it in order to split the whole character's storyline into successive narrative episodes. Such a partitioning depends on a threshold τ set by the user himself, depending on his specific information needs and preferences: τ corresponds to the maximal admissible distance between the most covering relationship vector in each narrative episode and any other relationship vector within this narrative episode. Such a threshold can be interpreted as the level of granularity desired when analyzing the story. We partition the whole set of scenes (storyline) into disjoint subsets of contiguous scenes (narrative episodes) by adapting a standard set covering model to this partitioning purpose.

As stated in (Daskin, 2011) in the context of optimal facility location, the set covering problem consists in locating a minimum number of facilities in a network so that all demand nodes are close enough (*i. e.* less than the maximum distance τ) to the nearest facility. In such a case, each network node, as a potential site for locating facilities, defines a subset of nodes distant from less than τ, and the general set covering problem consists in finding the minimum number of these subsets such that their union covers all nodes. Our narrative partitioning task can be formulated as a similar set covering

problem but with the following non-standard additional constraints:

- First, a constraint of temporal contiguity is put on the elements of the admissible subsets of scenes, so as to keep narrative episodes continuous over time.
- Second, in order to obtain a covering as close as possible to a partition, we minimize the overlap between the covering subsets instead of minimizing their number as in the standard formulation of the set covering problem.
- Despite this adapted objective, some relationship vectors at the boundaries between two consecutive narrative episodes may still belong to both of them: in this case, the covering is refined into a real partition by assigning the duplicated vector to the closest relationship state.

We now detail the way we adapt and solve the standard formulation of the set covering problem. First, the resolution of the set covering problem relies on the construction of an intermediate boolean matrix $\mathcal{A}$ from the distance matrix $\mathcal{D}$ and the threshold τ according to the following rule:

$$a_{ij} = \begin{cases} 1 & \text{if } d_{ij} \leqslant \tau \\ 0 & \text{otherwise} \end{cases}$$

Fig. 5.2 shows the boolean matrices $\mathcal{A}$ obtained after thresholding with $\tau := 1.0$ both matrices $\mathcal{D}$ shown on Fig. 5.1.

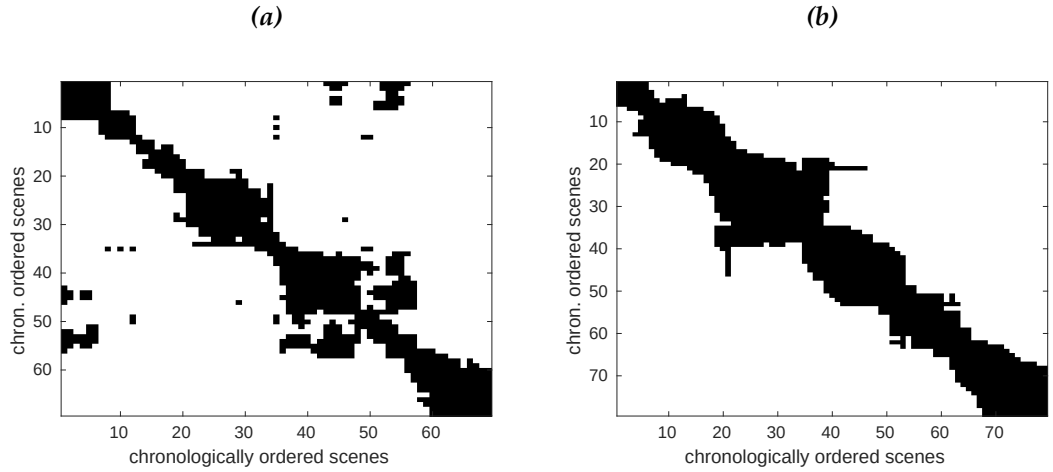


Figure 5.2 – *Boolean matrix $\mathcal{A}$ obtained after thresholding ($\tau := 1.0$) the matrix $\mathcal{D}$ containing the distances between Jaime Lannister's (5.2a) and Arya Stark's (5.2b) social neighborhoods in any pair of scenes (t, t') in the first five seasons of* Game of Thrones.

In each column (or, equivalently, row) j of the matrices $\mathcal{A}$ shown on Fig. 5.2, the set of values equal to 1 (in black) defines by construction a subset of scenes where the relationships of the character are close (*i. e.* under the threshold τ) to his relationships in the j^{th} scene. As can be seen by comparing Subfig. 5.2a with Subfig. 5.2b, both Jaime's and Arya's storylines look quite different from one another: while Arya's storyline is

linear, with successive social environments very distant from each other, Jaime's storyline is more *circular*: Jaime's social neighborhood around scenes 45–55 is quite similar to his relationships at the very beginning of the serial (scenes 1–10). Indeed, Jaime then comes back to King's Landing and is finally reunited with his siblings. Nonetheless, though semantically interesting, such social comebacks to previous relational states are not relevant for our partitioning purpose. We then further refine the matrix $\mathcal{A}$ of admissible subsets so that each subset contains only successive scenes: each narrative episode must be temporally continuous.

For this purpose, we set to 0 in each column j (resp. each row i) of the matrix $\mathcal{A}$ all the coefficients that are not related to a_{jj} (resp. a_{ii}) by a continuous sequence of ones. The resulting matrices, once filtered so that every admissible subset only contains contiguous scenes, are shown on Fig. 5.3, both for Jaime's (5.3a) and Arya's (5.3b) social environments.

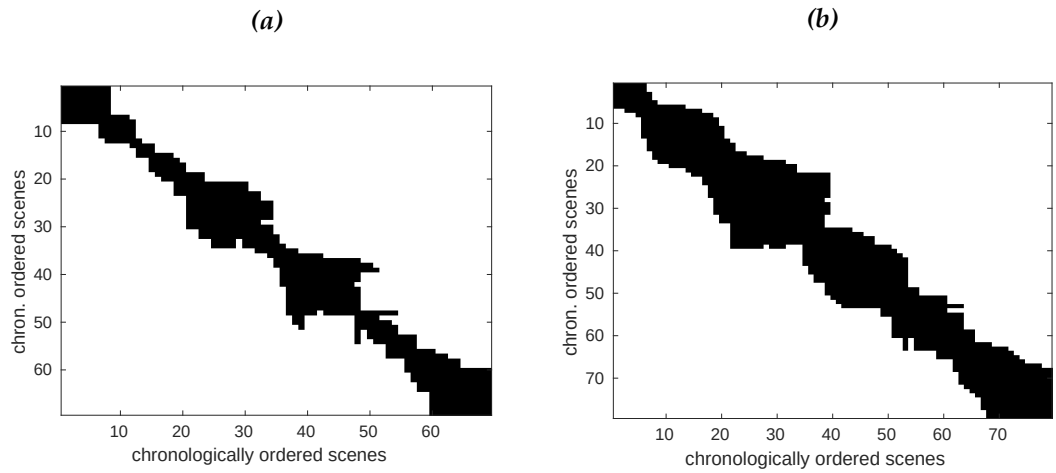


Figure 5.3 – *Filtered boolean matrix $\mathcal{A}$ only containing admissible subsets of contiguous scenes, for Jaime Lannister's (5.2a) and Arya Stark's (5.2b) social neighborhoods in any pair of scenes (t, t') in the first five seasons of* Game of Thrones.

The adapted set covering problem consists then in finding covering subsets with minimum overlap and is formulated as follows:

$$\text{(P4)} \begin{cases} \min \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_j \right) \\ \text{s.t.} \begin{cases} \sum_{j=1}^{n} a_{ij} x_i \geqslant 1 & i = 1, ..., n \\ x_i \in \{0, 1\} & i = 1, ..., n \end{cases} \end{cases}$$

where n is the total number of scenes, and each x_i $(i = 1, ..., n)$ defines an admissible subset which is set to 1 if selected, to 0 otherwise. The constraint $\sum_{j=1}^{n} a_{ij} x_i \geqslant 1$ $(i =$

$1, ..., n$) requires that each scene belongs to at least one of the selected subsets (covering constraint).

Fig. 5.4 shows the resulting partition of Jaime Lannister's (5.4a) and Arya Stark's (5.4b) distance matrix $\mathcal{D}$, with a granularity level $\tau = 1.0$. Each narrative episode is represented as a box containing a vertical line. This line corresponds to the scene in which the relationship state covers at best the narrative episode (x_j optimally set to 1). In such scenes, the relationship vector can be regarded as conveying the typical social environment of the character within the associated narrative episode.

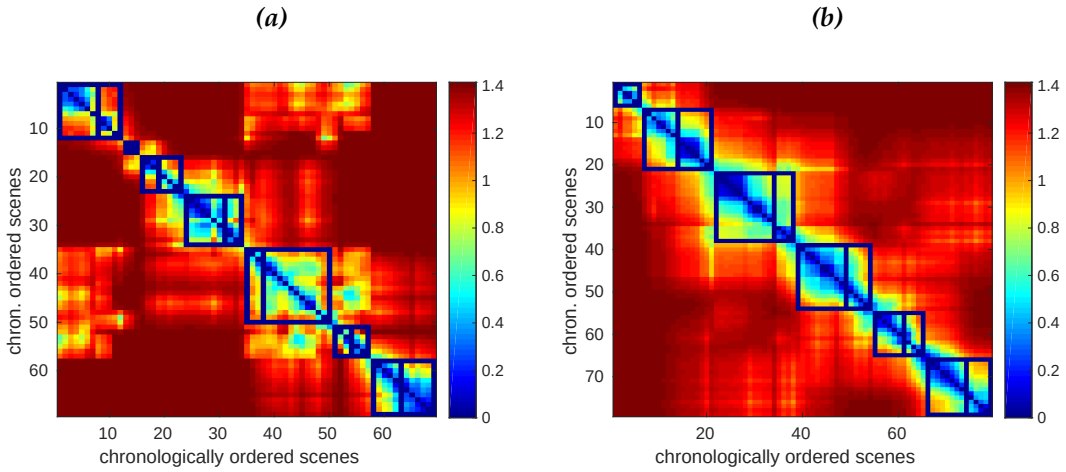


Figure 5.4 – *Partitioned matrix of distances between Jaime Lannister's (5.4a) and Arya Stark's (5.4b) social neighborhoods in any pair of scenes* (t, t') *in the first five seasons of* Game of Thrones.

For instance, the following representative vectors, as resulting from our partitioning method, provide us with Arya's social environment in the third and fourth narrative episodes (the components are re-arranged in decreasing order):

$$\mathbf{r_{34}} = \begin{pmatrix} \text{Tywin Lannister} & [0.82] \\ \text{Jaqen H'ghar} & [0.23] \\ \text{Hot Pie} & [0.21] \\ \text{Amory Lorch} & [0.21] \\ \vdots & \end{pmatrix} \quad \mathbf{r_{49}} = \begin{pmatrix} \text{Beric Dondarrion} & [0.54] \\ \text{Thoros} & [0.51] \\ \text{Anguy} & [0.51] \\ \text{Sandor Clegane} & [0.50] \\ \vdots & \end{pmatrix}$$

These two relationship vectors turn out to perfectly match two major developments in Arya's story: "Arya Stark captive in Harrenhal" ($\mathbf{r_{34}}$) and "Arya Stark and the Brotherhood" ($\mathbf{r_{49}}$). Similarly, the narrative episode mentioned above as "Jaime Lannister's journey in Dorne to rescue Myrcella" and captured by the relationship vector 5.1, corresponds to the vertical line in the rightmost block of Subfig. 5.4a. Appendix C provides representations of the matrix $\mathcal{D}$, before and after partitioning, for the five *Game of Thrones*' characters considered in the user study reported in Section 5.5.

5.3.3 Social relevance

For a given character, we define the *social relevance* $\mathbf{sr_i}$ of the i^{th} video sequence as the cosine similarity between the representative vector $\mathbf{r_t}$ of the character's relationships in the narrative episode to which the sequence belongs and the vector of relations the character is currently having within the i^{th} video sequence. As mentioned above, the representative vector $\mathbf{r_t}$ is derived from the smoothing method introduced in Section 4.4, whereas the components of the character's relationship vector within each sequence correspond to the interaction times between the character and every other character in the sequence considered, as estimated according to the basic heuristics described in Subsection 4.3.2. For a specific character, such a social relevance measure aims at discriminating the video sequences showing some of his typical relationships within each narrative episode of his storyline.

Nonetheless, social relevance remains too broad a criterion to be used on its own for discriminating relevant video sequences when building the summary, and the stylistic additional features described in Chapter 3 can help to isolate salient sequences among all those that are equally relevant from the social point of view.

5.4 Summarization algorithm

In this section, we describe the summarization framework we apply to generate our character-oriented summaries: we first introduce the weighting scheme we use to estimate the relevance of every candidate sequence; we then turn to the general formulation of the summarization problem as an integer quadratic mathematical program, and we finally detail the heuristic solution we introduce for solving large instances of the summarization problem.

5.4.1 Relevance weighting scheme

In Chapter 3, we introduced two saliency-oriented mid-level features, along with the associated extraction algorithms: shot size and musicality. Once measured, these two features can be averaged for each video sequence candidate for later insertion in the character-oriented summary, resulting for the i^{th} sequence in two quantities $\mathbf{ss_i}$ (average shot size) and $\mathbf{m_i}$ (average musicality). Both stylistic features are then combined with social relevance (denoted $\mathbf{sr}$) according to the following weighting scheme, resulting in a single measure of relevance p_i of the i^{th} candidate sequence for building the summary of a specific character's storyline:

$$p_i = \lambda_1 \mathbf{sr_i} + \lambda_2 \mathbf{ss_i} + \lambda_3 \mathbf{m_i} \tag{5.2}$$

Social relevance and average shot size, by construction, range from 0 to 1. We therefore *min-max* normalize musicality to get values between 0 and 1 for this feature. We set

the vector of weights to $\boldsymbol{\lambda} := (0.25, 1.0, 1.0)^\intercal$ to obtain a reasonable trade-off between the content and style-based features.

As stated in Subsection 3.2.4, we choose short (5–15 seconds long) *logical story units* as the basic candidate units for later insertion in our character-oriented summary. We now describe how we build the summary of the storyline associated to a specific character by iteratively selecting optimal candidate LSUs, after weighting their relevance.

5.4.2 Summarization problem

Character-oriented summaries aim at reflecting the dynamics of a character's storyline. Once isolated by applying the segmentation method described in Subsection 5.3.2, each narrative episode of a specific character's storyline should be equally reflected, whatever its duration, as a major development in his story. We therefore build the summary step-by-step to reflect the natural segmentation of the storyline into narrative episodes, possibly of variable duration.

Our algorithm for constructing the character-oriented summary takes two inputs, set by the user himself depending on his information needs and subjective preferences: the level of granularity τ for analyzing the storyline in narrative episodes; and the maximum time T, expressed in seconds, devoted in the summary to each narrative episode. Each narrative episode, once isolated, consists of a subset of scenes, containing n candidate LSUs, each i^{th} LSU being weighted according to the global relevance score p_i introduced in Equation 5.2.

The building of the summary can then be regarded as a task with two joint objectives and a length constraint: the summary of the considered narrative episode must not exceed the duration T and aims at containing not only relevant sequences, but also sequences that remain as diverse as possible, in order to minimize redundancy. In (McDonald, 2007), the author shows that such a summarization problem can be formulated as the following quadratic knapsack problem, with two joint objectives and a length constraint:

$$\text{(P5)} \begin{cases} \max f(\mathbf{x}) = \left(\sum_{i=1}^{n} p_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_i x_j \right) \\ \text{s.t.} \begin{cases} \sum_{i=1}^{n} w_i x_i \leqslant T & i = 1, ..., n \\ x_i \in \{0, 1\} & i = 1, ..., n \end{cases} \end{cases} \tag{5.3}$$

where x_i is a binary variable set to 1 if the i^{th} LSU is inserted in the summary, and to 0 otherwise; p_i denotes the relevance, as computed according to Equation 5.2, of the i^{th} LSU; w_i is the duration, expressed in seconds, of the i^{th} LSU; T is the maximum time devoted in the summary to the narrative episode containing the current subset of n LSUs; and finally, d_{ij} is a measure of difference between the i^{th} and j^{th} LSUs: it is defined as

the normalized euclidean distance between the vectors of direct relationships in the i^{th} and j^{th} LSUs, as defined in Subsection 5.3.3. The maximization of the first, linear part $\sum_{i=1}^{n} p_i x_i$ of the objective function $f(\mathbf{x})$ aims at selecting relevant sequences, both stylistically salient and socially representative of the character's relationships in the current narrative episode. The maximization of the second, quadratic part $\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} x_i x_j$ of the objective conveys the goal of social non-redundancy: once introduced in the objective function, the coefficients d_{ij} aim at generating a summary that provides us with an overview of the full range of the character's relationships at this point of the story, instead of focusing on a single representative relationship shown in several redundant sequences.

5.4.3 Heuristic solution

As stated in (McDonald, 2007) and (Gillick et Favre, 2009), the formulation 5.3 of the summarization problem can be tricky to solve exactly for large instances, even when linearizing the quadratic part of the objective function, and heuristic methods provide us with more scalable, though possibly sub-optimal, resolution techniques. However, in (McDonald, 2007), the author underlines the limitations of Maximal Margin Relevance (MMR) based algorithms (Carbonell et Goldstein, 1998) to achieve the double objective of relevance and diversity when building summaries: if selected iteratively without taking their length into account, the already selected sequences may be too long and prevent us from choosing additional sequences to improve the objective function.

As a possible alternative to the dynamic programming-based algorithm detailed in (McDonald, 2007), we introduce here another greedy heuristic for iteratively selecting LSUs in a more optimal way than the MMR algorithm does. Our approach is summarized in Algorithm 2: it generalizes to the quadratic case the usual greedy heuristic used to solve the linear knapsack problem (Fayard et Plateau, 1982; Balas et Zemel, 1980).

The summary S of the currently considered narrative episode consists of the indices of the selected LSUs and is built iteratively from the whole set L containing the indices of the candidate LSUs. At each iteration, we choose the LSU with the maximum relevance/duration ratio (line 4), whose duration does not exceed the remaining time still available and which does not overlap with any of the previously selected LSUs (conditions line 5): some candidate LSUs may share some shots and partially overlap. Such a heuristic, by focusing on sequences with optimal ratios (p_i/w_i) tends to select short LSUs, as long as they are relevant enough.

Furthermore, the additional objective of non-redundancy is taken into account when iteratively updating the relevance of the remaining candidate LSUs (line 9): the updated relevance not only comes from the LSU intrinsic relevance (p_i) but also from his distance

Algorithm 2 LSUs SELECTION

Require: $\mathbf{p}, \mathbf{w} \in \mathbb{R}_+^n$, $T \in \mathbb{R}_+$, $(d_{ij})_{i,j=1,...n} \in \mathbb{R}_+$
$L \leftarrow \{1, ..., n\}$
$S \leftarrow \emptyset$
while $L \neq \emptyset$ **and** $T > 0$ **do**
  $s \leftarrow \underset{i \in L}{\text{argmax}}(p_i / w_i)$
  if $w_s \leqslant T$ **and** non_overlap(s, S) **then**
    $S \leftarrow S \cup \{s\}$
    $T \leftarrow T - w_s$
    for $i \leftarrow 1$ to n **do**
      $p_i \leftarrow p_i + 2d_{is}$
    end for
  end if
  $L \leftarrow L - \{s\}$
end while
return S

$(2d_{is})$ to the previously selected LSUs. The detail of the updating formula comes from the re-writing of the objective function after each new sequence is selected. Assuming for instance and without loss of generality that after the very first iteration of the algorithm, the first LSU is selected, the objective function can be re-written as:

$$\begin{aligned} f(\mathbf{x}) &= (p_1 + d_{11}) + \sum_{i=2}^{n} p_i x_i + \sum_{i=2}^{n}\sum_{j=2}^{n} d_{ij} x_i x_j + \sum_{j=2}^{n} d_{1j} x_j + \sum_{i=2}^{n} d_{i1} x_i \\ &= (p_1 + d_{11}) + \sum_{i=2}^{n} p_i x_i + \sum_{i=2}^{n}\sum_{j=2}^{n} d_{ij} x_i x_j + \sum_{i=2}^{n} (d_{i1} + d_{1i}) x_i \\ &= (p_1 + d_{11}) + \sum_{i=2}^{n} p_i x_i + \sum_{i=2}^{n}\sum_{j=2}^{n} d_{ij} x_i x_j + \sum_{i=2}^{n} 2d_{i1} x_i \\ &= (p_1 + d_{11}) + \left(\sum_{i=2}^{n} (p_i + 2d_{i1}) x_i + \sum_{i=2}^{n}\sum_{j=2}^{n} d_{ij} x_i x_j \right) \end{aligned} \tag{5.4}$$

where the right-hand part of the last equation is the objective function of the summarization problem 5.3 for the remaining, not yet selected, LSUs, again formulated as the objective function of a quadratic knapsack problem, but with an updated vector of relevance values: after one LSU is selected, the relevance of the remaining sequences combines both their intrinsic relevance p_i and their distance to the previously selected LSUs.

The heuristic method implemented by Algorithm 2 turns out to be fast, even when applied to large instances of the problem, and to provide solutions very close to the optimal solutions. For each narrative episode, a summary is built until the time duration limit T is reached, and the final character-oriented summary is made of the concatena-

tion of all the LSUs, chronologically re-ordered, selected in every narrative episode.

5.5 Experiments and results

In order to evaluate our method for automatically generating character-oriented summaries of TV serials, we performed a large-scale user study in a real case scenario. In this section, we first describe the user study we performed; we then explain the types of summaries the participants were asked to rank, along with the evaluation protocol; finally, we detail the results we obtained.

5.5.1 User study

A few weeks before the new, sixth season of the popular TV serial *Game of Thrones* (denoted hereafter GOT) was released, people, mainly students and staff of our university, were asked to answer a questionnaire, both in order to collect various data about their TV series viewing habits and to evaluate automatic summaries centered on five characters of GOT. A total of 187 subjects took part in the questionnaire, that we designed according to the principles detailed in (Ethis et Fabiani, 2004), with 52.7% female, and 47.3% male participants. The population was quite young: 21.14 years old in average, $\pm$2.67 years. The first 37 participants answered the questionnaire in controlled conditions, in the same room and with possible help, while the remaining others answered online without help. Every answer was collected within a period of 8 days. Being familiar with GOT, if recommended, was not mandatory to answer the questionnaire. 27% of the people we polled had actually never watched GOT when answering the questionnaire, but 56% of them had seen all first five seasons. As shown on Fig. 5.5, more than half of the GOT's viewers we polled in our study feel the need to remember the plot of the past seasons when a new one is about to be released.

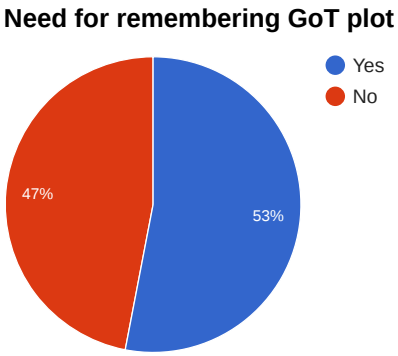


Figure 5.5 – *Do you feel the need to remember the plot of* Game of Thrones *previous seasons?*

Surprisingly, such a proportion is not as important as for TV series in general (nearly 60%, as shown on Fig. 1.5): the plot of GoT, especially complex with multiple parallel storylines, was expected to cause such a need at a higher degree. Nonetheless, this TV serial is popular enough to provoke many discussions, probably able to partially keep the memory of the plot. As shown on Fig. 5.6, in the special case of GoT, such discussions are even preferred, without excluding them, to textual or video summaries to obtain the revival effect we target: 57.6% of the people we polled discuss with friends to remember the plot of GoT, 32.9% read textual summaries and 34.1% watch video summaries, whereas Fig. 1.6 shows that these three channels of information are roughly equally used for TV series in general.

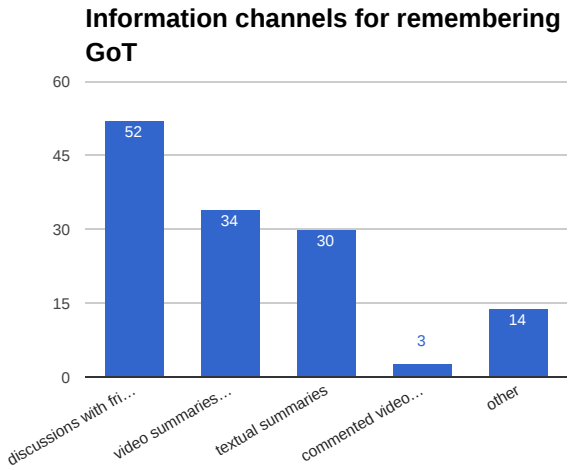


Figure 5.6 – *Before viewing a new season of* Game of Thrones, *which information channel(s) do you use for remembering the plot of the previous seasons? (multiple answers allowed)*

As shown on Fig. 5.7, about 60% of the people we polled had last watched GoT more than six months ago, and were in the typical use case we detailed in Chapter 1.

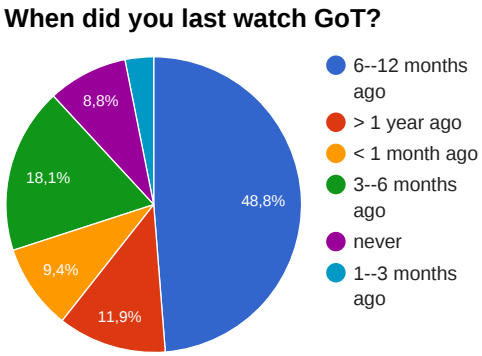


Figure 5.7 – *When did you last watch* Game of Thrones?

Such a proportion even increases, amounting to 64.2%, among the people who had watched all first five seasons of GOT.

5.5.2 Summaries for evaluation

After answering general questions about their TV series viewing habits, the participants were asked to evaluate summaries of GOT centered on the storylines of five characters.

TV series data

By focusing on such a popular TV serial, we were quite certain that most participants would have watched it; moreover, with GOT, we were at the time of the study in the typical use case we want to target (see Chapter 1), with a new season about to be released. Unfortunately, the other two TV serials of our corpus did not fulfill all these requirements: first, they are not as popular as *Game of Thrones*, and by focusing on these TV serials, we could not have performed a user study at such a large scale. Moreover, at the time of the study, *Breaking bad* was completed, and the last season of *House of Cards* had recently been released. So, with both these series, we were not in the typical use case that motivates the need for TV serial summaries.

Our character-oriented summaries are generated from partially annotated data. The corpus we used covers the whole set of the first five seasons of GOT (50 one-hour episodes). As stated in Chapter 4, we manually inserted the scene boundaries within each episode and we labeled every subtitle according to the corresponding speaker, as a basis for estimating verbal interactions between characters and building the dynamic social network of interacting speakers. Though it would have been possible to automatically perform both tasks, scene boundaries detection and speaker detection, the second one either in a supervised (speaker recognition) or in an unsupervised way (speaker diarization), we decided to hand-label these data: as mentioned in Subsection 3.5.7, we wanted to do a relatively large-scale study and could not afford to show the viewers in a limited time of 1 hour both fully and partially automatic summaries to measure the impact of the errors made at the speaker recognition/diarization level; moreover, GOT contains many speakers, even when focusing on the major ones, often speaking in adverse conditions (background music, sound effects) and the error rates obtained when automatically performing speaker diarization (detailed in Chapter 3) were too high to serve as a reliable basis for building with confidence the dynamic social network of interacting speakers.

Summary generation

In order to get generalizable user feedback on our summarization framework, we asked the users to evaluate the summaries of five characters' storylines.

Our criteria to select the five characters were the following:

- These characters were still involved in the narrative at the end of the last season of GOT, and were consequently likely to play a role in the incoming one.

- They all are important enough to have evolved at some point of the plot in their own storylines.
- Their story is likely to be complex enough to require a summary before viewing the next season.

These five characters are: Arya Stark, Daenerys Targaryen, Jaime Lannister, Sansa Stark, and Theon Greyjoy.

The summaries cover the storylines of these five characters over all the first five seasons of GOT. Such long-term summaries are expected to capture the whole dynamics of a character's storyline when introducing to the next season, rather than only focusing on the very last events he happened to experience during the last season. The more a plot is advanced, the more such long-term summaries are probably needed, especially when the plot is complex. We automatically created three summaries for each character:

- First, a full summary (denoted **full**), resulting from the application of the algorithm we described in Section 5.4 and designed so as to be sensitive to both the content and the style of the narrative. This first summary depends on two user-dependent parameters: the granularity level τ used for segmenting the storyline into narrative episodes, and the time T devoted in the summary to every resulting narrative episode. In order to keep the summary duration into reasonable boundaries, we chose the following parameter setting: $\tau = 1.0$ and $T = 25$ seconds.
- Second, a style-based summary (denoted **sty**), only built upon the stylistic features we described in Sections 3.3 (shot size) and 3.4 (background music). Social relevance is then ignored, and the weights of the weighted sum in Equation 5.2 are set to $\boldsymbol{\lambda} := (0.0, 1.0, 1.0)^\intercal$. In this case, there is no pre-segmentation step of the storyline into narrative episodes based on the dynamic social network of interacting speakers. As a result, the candidate LSUs are not selected among the separate subsets resulting from the segmentation step; instead they are considered as a whole single set of candidate sequences, weighted according to their average shot size and musicality, and finally selected by applying Algorithm 2 with all coefficients d_{ij} set to 0, until the resulting style-based summary has roughly the same duration as the full summary.
- Third, a baseline, semi-random summary (denoted **bsl**) is obtained as follows: some non-overlapping LSUs where the considered character is verbally active are first randomly selected until reaching a duration comparable to the duration of the first two kinds of summaries; the selected LSUs are then re-ordered chronologically when inserted in the summary.

Summary features

The main properties of the resulting three types of summaries are reported in Table 5.1 for each of the five considered characters. For each character and each type of summary, the number of candidate LSUs is mentioned, along with their average duration in seconds. The same properties are reported for the selected LSUs inserted in the summary. The total duration of the resulting summary, expressed in seconds, is mentioned in the

seventh column. Finally, the compression rate is mentioned in the last one, computed as the ratio between the total duration of all scenes in which the character is verbally active and the summary duration.

Table 5.1 – *Properties of the three types of summary generated for each character's storyline during the first five seasons of* GOT*: number and average duration of candidate and selected* LSUS*, summary duration and compression rate.*

character	summary	LSUs				dur.	comp. rate
		candidate		selected			
		#	dur.	#	dur.		
Arya	**full**	2,156	10.4	24	5.7	137.7	80.2
	sty	2,180	10.4	24	6.1	145.3	76.0
	bsl	2,180	10.4	14	10.4	145.1	76.1
Daenerys	**full**	1,171	10.6	15	6.3	93.8	139.0
	sty	1,185	10.6	16	6.0	96.6	135.0
	bsl	1,185	10.6	10	9.5	95.5	136.9
Jaime	**full**	962	11.0	25	6.4	153.4	71.2
	sty	963	11.0	25	6.9	172.4	65.9
	bsl	963	11.0	15	11.1	167	68.0
Sansa	**full**	888	11.0	24	5.8	139.6	91.4
	sty	892	11.0	24	6.1	146.1	87.4
	bsl	892	11.0	15	9.7	146.2	87.4
Theon	**full**	650	10.9	16	6.0	95.7	81.6
	sty	655	10.9	15	6.4	96.0	81.3
	bsl	655	10.9	9	10.9	98.3	79.4

As can be seen, the number of candidate LSUs differs from one character to the other: for each character, the only LSUs considered are those where he/she is verbally active in order to center the summary on this specific character. Moreover, the style-based and baseline summaries rely on slightly more candidate LSUs than the full ones: a few scenes only containing LSUs with isolated utterances of the character with no hypothesized interlocutor were discarded when building the full summaries, but not when constructing the other two types. By definition, social relevance can not apply to these LSUs hypothesized as soliloquies, but music and shot size may nonetheless make them stylistically salient.

The final summaries turn out to be quite short, ranging from 1:30 to 2:50 minutes, resulting in very high compression rates: the whole story of a character during 50 one-hour episodes is summarized in about two minutes. When summarizing the storyline of important characters, like Daenerys Targaryen, the compression rate may be much higher than when summarizing the storylines of characters that are not as important. The total time of the summary is actually dependent on the number of narrative episodes resulting from the segmentation of the storyline based on the character's evolving social network: characters with fast-evolving social environments, going through more narrative episodes, may therefore need longer summaries than possibly

more important characters involved in fewer narrative episodes. For instance, Fig. C.2 shows that Daenerys' storyline can be split in only 4 narrative episodes, because of her slow-evolving social environment. In contrast, though Jaime appears in fewer scenes than Daenerys ($\simeq$ 70 instead of $\simeq$ 85), the social analysis of his storyline at the same granularity level requires many more narrative episodes (7 instead of 4, as can be seen on Fig. C.3), resulting in longer summaries with lower compression rates.

The duration of the full summary is sometimes not as long as the style-based and baseline ones. When building full summaries, LSUs are selected separately in each narrative episode, until the limit of 25 seconds is reached for each one. This may result in a cumulative loss of a few seconds with respect to the total time available (25 seconds $\times$ number of narrative episodes). For both other types of summaries, with LSUs selected from a single global set, the loss is usually not as important and the global time limit is nearly reached. Not surprisingly, the heuristic used when building the full and style-based summaries by applying Algorithm 2 results in summaries consisting of shorter sequences than the baseline summary: while the candidate LSUs last a bit more than 10 seconds in average (fourth column in Table 5.1), the duration of the selected ones (sixth column) in the **full** and **sty** summaries, based on an optimal ratio between relevance and duration, is very close to the lower bound of 5 seconds put on the duration of the candidate LSUs (see Subsection 5.4.1). On the opposite, the sequences inserted in the **bsl** summaries are almost twice as long and very close to the average duration of the candidate ones.

Finally, the three summaries may overlap. Table 5.2 reports for each of the five considered characters the overlapping time, expressed in %, between the three summary types.

***Table 5.2** – Overlapping time (in %) between the three summaries for each considered character.*

Summaries	*Character*				
	Arya St.	Daenerys	Jaime L.	Sansa St.	Theon Gr.
bsl / full	4.85	3.12	1.70	2.03	4.85
bsl / sty	12.62	9.80	0.75	9.11	6.28
full / sty	30.26	32.61	32.80	35.54	36.07

The overlapping time between the **full** and **sty** summaries, ranging from 30% to 36%, is remarkably constant, because of the stylistic features the two summarization algorithms share; as expected, the overlapping time between the **full** and **sty** summaries on the one hand, and the baseline summary **bsl** on the other, is not as important and ranges from 2% to 10%, depending on the considered character.

5.5.3 Evaluation protocol

The users were asked to rank, for each character, the three summaries according to the two usual criteria used in subjective evaluation of summaries, *informativeness* and

enjoyability (Truong et Venkatesh, 2007), but reformulated as follows according to the specific use case we target:

1. Which of these three summaries reminds you the most the character's story?

2. Which of these three summaries makes you the most wanting to know what happens next to the character?

The same questions were asked for their last choice, resulting in a full ranking of the three summaries for each character. In addition, the participants were asked to motivate in a few words their ranking. Moreover, answering the ranking questions was not mandatory, if the participants were too unsure.

No restriction was put on the number of possible viewings of the three summaries: a passive, first viewing of the summaries was actually expected to be needed to help the viewer remember the main steps of the character's storyline; a second, informed viewing was then expected to be possibly needed to finely compare the summaries according to the two criteria (hereafter referred to as "*best as recap?*" and "*best as trailer?*", respectively). About 25% or the participants in average needed several viewings to rank the three summaries according to the two criteria.

5.5.4 Results

For each character's storyline, the best summaries according to those of the participants who had watched all past five seasons of GoT are reported in Table 5.3, both as a proportion (denoted "%") and a number (denoted "#") of participants.

Table 5.3 – *For each character's storyline, best summary according to the participants who had watched all past five seasons of* GoT.

character	Result	*best as recap?*			*best as trailer?*		
		full	**sty**	**bsl**	**full**	**sty**	**bsl**
Arya	%	**70.9**	9.3	19.8	**57.1**	16.7	26.2
	#	***61***	*8*	*17*	***48***	*14*	*22*
Daenerys	%	**35.8**	32.8	31.3	18.2	**47.0**	34.8
	#	***24***	*22*	*21*	*12*	***31***	*23*
Jaime	%	**41.5**	40.0	18.5	35.9	**43.8**	20.3
	#	***27***	*26*	*12*	*23*	***28***	*13*
Sansa	%	**47.7**	33.8	18.5	**58.5**	20.0	21.5
	#	***31***	*22*	*12*	***38***	*13*	*14*
Theon	%	15.6	**45.3**	39.1	14.3	**57.1**	28.6
	#	*10*	***29***	*25*	*9*	***36***	*18*
all	%	**42.3**	32.2	25.4	36.8	**36.9**	26.3

As can be seen in Table 5.3, baseline summaries never obtained a majority vote, whatever the ranking criterion: for three of these summaries (Arya, Jaime, Sansa), the

scores remain quite low. In some cases, according to the feedback we got, participants chose the baseline summary because of the length of the sequences (about 10 seconds, twice as much as in the other summary types), perceived as more appropriate to fully understand and remember the selected sequences. However, setting the lower bound of the admissible LSUs to a higher value would first have resulted in too long summaries: many users reported during our study that 2-3 minutes was the maximum duration they could tolerate for summaries of TV series[1]. Furthermore, it would have made the user's feedback trickier to interpret: our method aims at showing both socially relevant and stylistically salient verbal interactions between characters, but is not sensitive to their semantic content. Long sequences may unintentionally introduce into the user's perception accidental semantic effects that would upset the subjective evaluation of our summarization framework.

Nonetheless, concise summaries with short, socially diverse sequences extracted from every narrative episode were globally well perceived. For 4 out the 5 characters targeted, the full summary was selected as the most efficient recap, in some cases by far: Arya's story full summary for instance was considered as the best recap by 70.9% of the participants. Even when including in the evaluation process all participants, possibly unaware of every development of Arya's storyline, the full summary is preferred by a vast majority as the best recap as shown on Fig. 5.8.

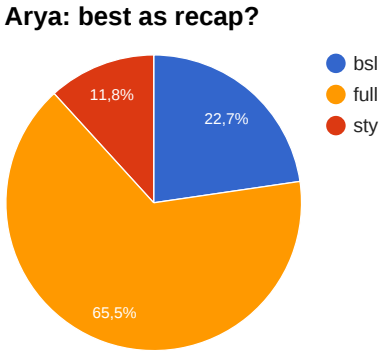


Figure 5.8 – *Subjective evaluation of Arya's storyline summaries as effective recaps (all participants).*

Fig. 5.9 shows the distribution of the selected LSUs in both style-based (5.9a) and full (5.9b) summaries of Arya's storyline.

As can be seen, the last narrative episode (last blue box from left to right on both figures), is not represented in the style-based summary, in contrast to the full summary. As a result, the last, fifth season of *Game of Thrones* is not represented in Arya's style-based summary, which was perceived as "incomplete" by many participants. Moreover, stylistic saliency is likely to be underrepresented in short narrative episodes, as

[1] As stated in (Ethis, 2006), the perception of time in fictional works is quite subjective.

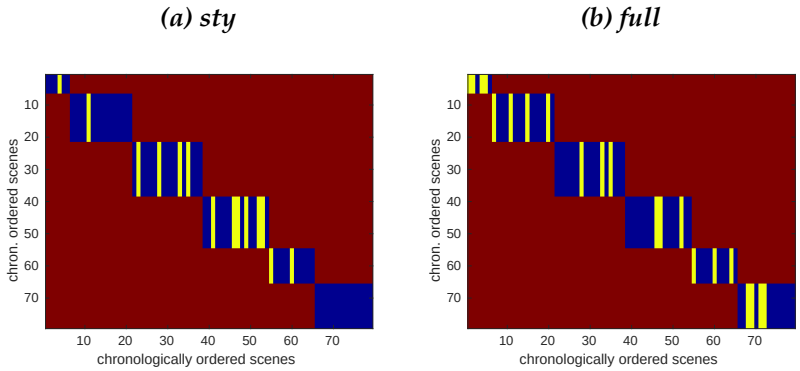


Figure 5.9 – *Distribution of the selected* LSUs *(yellow vertical lines) over the narrative episodes (blue boxes) for Arya Stark's storyline summaries: style-based (**sty**, 5.9a) and full (**full**, 5.9b).*

the first one (first blue box from left to right on both figures) in Arya's storyline, resulting in incomplete summaries unable to capture the whole dynamics of the character's storyline.

Furthermore, for 4 out of the 5 characters (Arya, Daenerys, Jaime, Theon), the full summaries obtain higher scores when judged as recaps than when judged as trailers. In some cases, the difference is impressive: whereas 35.8% of the participants who had watched all seasons of GoT rank the full summary of Daenerys' story as the best recap, they are only 18.2% to rank it as the best trailer. Even when taking into account the votes of all participants, and not only the votes of those who had watched all first five seasons of GoT, the difference of ranking resulting from the type of criterion applied is striking, as shown on 5.10.

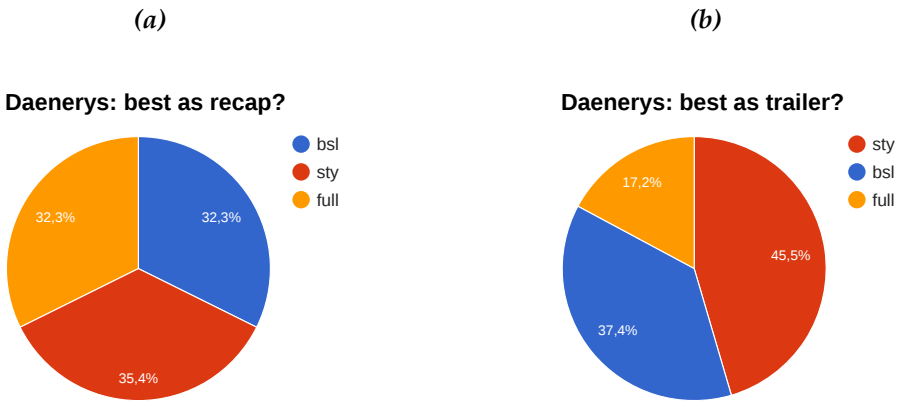


Figure 5.10 – *Subjective evaluation of Daenerys' storyline summaries: as the best recap (5.10a), and as the best trailer (5.10b) (all participants).*

Such a difference of ranking when switching from the "recap" to the "trailer" criterion globally benefits the style-based summaries, more appreciated as trailers than as recaps for 4 characters out of 5 (Arya, Daenerys, Jaime, Theon). As can be seen for

instance on Fig. 5.10, most of the votes are transferred from the full to the style-based summary when switching from criterion. For 3 characters (Daenerys, Jaime, Theon), such style-based summaries even obtain a majority vote according to the "trailer" criterion.

However, some of the results we obtained were sometimes unexpected. First, as can be seen from Subfig. 5.10a, the three summaries of Daenerys' storyline obtain roughly similar scores when evaluated as recaps, without clear advantage for the full summary: Daenerys is a key-character of GOT, often named "Mother of Dragons" from the fact she owes three dragons. Many participants, judging from the short explanations they gave to motivate their ranking, turned out to focus on this aspect of Daenerys to assess the summaries: absent from the full summary, Daenerys' dragons are heard in the style-based one, and seen in the baseline one. Such a criterion was sometimes used to discard the full summary, though being the only one that captured the very last narrative episode of her storyline, *i. e.* her crucial meeting with Tyrion Lannister (last, small blue box from left to right on both parts of Fig. 5.11).

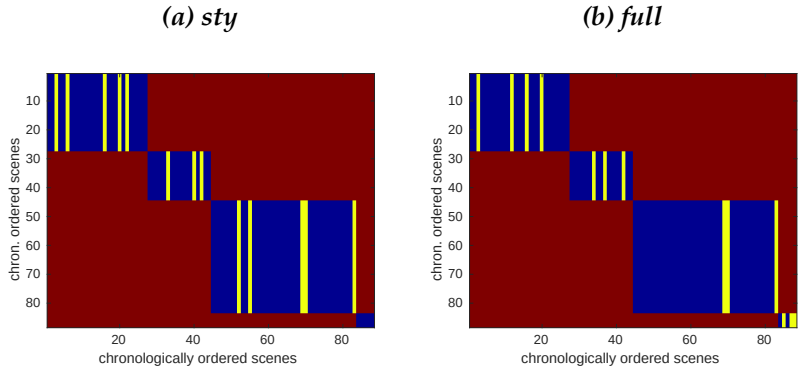


Figure 5.11 – *Distribution of the selected* LSUs *(yellow vertical lines) over the narrative episodes (blue boxes) for Daenerys Targaryen's storyline summaries: style-based (**sty**, 5.11a) and full (**full**, 5.11b).*

The scores obtained by Theon's summary were also surprising, with quite low scores for the full summary, probably penalized by a baseline summary rather semantically consistent and convincing, though somehow incomplete: two major narrative episodes, the "Fall of Winterfell" and the "Final Reunion with Sansa" are missing in the baseline summary, but well captured by the full summary (respectively in the 3rd and 4th boxes shown on Fig. 5.12).

Nonetheless, the results we obtained globally strengthen our "plot modeling" approach when it comes to summarize the dynamics of a character's storyline over dozens of episodes. At such narrative scales, relying on only stylistic patterns remains hazardous, and may lead to miss important developments in the plot, especially when occurring in short narrative episodes. In contrast, our SNA-based approach results in summaries covering enough to act on the viewers as effective recaps.

When it comes to emotionally re-engage the viewers in the plot, as trailers intend to, stylistic features, in addition to content-related one, are valuable. The two mid-

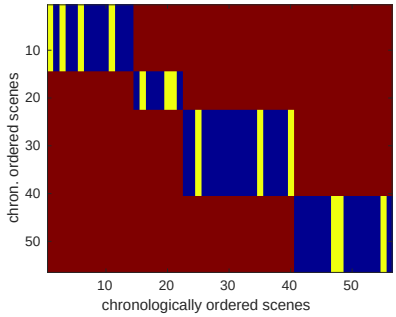


Figure 5.12 – *Distribution of the selected* LSUs *(yellow vertical lines) over the narrative episodes (blue boxes) for Theon Greyjoy's storyline full summary.*

level features we used to isolate re-engaging sequences, *i. e.* shot size and background music, turned out to effectively support such a trailer effect, resulting in summaries able to make the viewers wanting to know "what comes next" in the narrative.

5.6 Conclusion

In this chapter, we described and evaluated a way to automatically generate character-oriented summaries of TV serials.

We first described a method for modeling each character's storyline from the smoothed dynamic social network we introduced in Chapter 4, before detailing a weighting scheme for estimating the relevance of each candidate sequence. On the one hand, the relevance of each candidate sequence, based on a step of pre-segmentation of the character's storyline into narrative episodes, is expressed in terms of social relevance. The summary is then designed so as to focus on the most typical relationships of the character at each step of his/her storyline. On the other hand, we also expressed the relevance of each candidate sequence in terms of stylistic saliency. We specifically focused on shot size and background music, as mid-level features commonly used by filmmakers to emphasize the importance of some specific sequences.

We evaluated our summarization framework by performing a large-scale user study in a real case scenario by focusing on the popular TV serial *Game of Thrones* and obtained promising results. The social network perspective we introduced results in content-covering summaries that the viewers perceived as effective recaps of the complex plot of *Game of Thrones*; in addition, the use of background music and shot size for supporting the engaging, revival effect expected from such summaries turned out to be relevant to the users we polled.

Furthermore, our character-oriented summaries benefit from the empathetic relationship the viewer is likely to have with his/her favorite character(s): the cold-start phenomenon that the season trendlines of *IMDb* ratings exhibit for instance on Fig. 1.8 not only depends on the viewer's cognitive disengagement, but probably also on an

emotional disaffection that character-oriented summaries can handle properly.

Chapter 6

Conclusion and perspectives

Contents

6.1 Conclusion

TV series became increasingly popular these past ten years. However, the narrative continuity of the most popular TV series directly conflicts with modern viewing conditions, which turn out to be highly discontinuous. Such a contradiction results in common information needs that provide video summarization with novel and realistic use-case scenarios. Long restricted to trailer generation, automatic movie summarization finds with TV series a high-profile, high-profit and well-defined application field.

In Chapter 1, we motivated from the user's perspective the need for summaries of TV serials, *i. e.* TV series with continuous stories. Once released, new seasons of TV series are usually being watched over short periods of time, and viewers are likely to have to some extent forgotten the plot. Moreover, such waiting periods usually result in emotional disengagement, and the summaries of the past seasons are expected to revive the plot both from cognitive and affective points of view.

Automatic generation of extractive summaries to help remember such long-term stories raises novel and challenging issues in the video summarization field: by definition, single episodes of TV serials are not independent from each other, and can definitely not be regarded as narratively self-sufficient. At the episode scale, the saliency-oriented summarization frameworks described in Chapter 2 can still capture meaningful information: salient sequences detected from low-level features are likely to accidentally carry in addition significant information. As a consequence, the resulting summaries may seem not only engaging, but also content-covering enough to be evaluated according to the *informativeness* criterion. Nevertheless, at the much larger scale of dozens of consecutive episodes, saliency-oriented approaches prove to be far from sufficient: summaries only based on saliency markers were usually perceived by the viewers we polled in our user study (Chapter 5) as incompletely covering the source content, and content-oriented approaches are more appropriate to this situation.

Following research directions initiated in literary studies, we made use in Chapter 4 of *social network analysis* to model the dynamics of TV serial plots for content coverage purposes. Once speakers are detected (Section 3.5), the way they interact with one another can be estimated (Section 4.3), resulting in an evolving social network that properly reflects the social dynamics of the TV serial (Section 4.4). Based on such a dynamic conversational network, narrative units can be detected, either as dynamic communities of characters interacting at some point of the story, either in the storyline associated to a specific character (Section 5.3): content-covering summaries should then reflect equally every narrative episode, by showing for the considered character the full range of his/her relationships within each of these narrative developments. As reported in Section 5.5, such summaries were globally perceived as effective synopses for viewers that used to be at some point familiar with the source content.

6.2 Perspectives

6.2.1 Subtasks

Multi-modal speaker diarization

The multi-modal speaker diarization framework we introduced in Subsection 3.5.4 relies on late fusion between acoustic and visual partitions of the utterances contained in short dialogue sequences with two speakers. In order to detect recurring speakers within larger sequences, like *logical story units*, earlier fusion of acoustic and visual features could be performed by co-clustering both similar shots and similar utterances, with an additional objective of temporal alignment of the resulting shot and utterance clusters: the utterances covered by the same recurring shot are likely to originate in the same speaker, whereas the utterances originating in the same speaker might match several sets of recurring shots. The first results we obtained by performing such a co-clustering of shots and utterances by solving a single optimization problem with two joint objectives are promising, but not scalable enough to large sets of shots/utterances. Heuristic resolution techniques are likely to be more effective than the exact resolution approach we adopted in the first place.

Narrative smoothing

The technique we introduced in Section 4.4 to construct the dynamic social network of characters interacting in TV serials should generalize well to any other situation where possibly parallel interactions are sequentially monitored.The first obvious extension is to other narrative media, for instance literary works, where the information channel is sequential. But sequentiality may also result from the fact that the whole set of interactions if not fully observed at any moment. Some individuals for instance may punctually interact on a certain social medium, but may keep on interacting in a latent way on other media. We are then left with processing such possibly unobserved interactions when building the dynamic network of individuals interacting on the considered social medium. The conservative aspect of narrative smoothing may tackle conveniently such issues.

6.2.2 Plot modeling

Causal plot modeling

The way we partitioned in Section 5.3 a character's storyline into narrative episodes aimed at capturing views of his successive typical social environments. In doing so, we were implicitly interested in capturing what was "normal" at every narrative stage, rather than detecting breakpoints in the story. A further step towards plot modeling would be to capture in addition the punctual events that destabilize the character's social environment, or more generally the social structures in the story. Focusing on the narrative breakpoints at the boundaries between successive narrative episodes would

allow to capture what (Tsai et al., 2013) denote as "causal relationships" between characters. Such causal plot modeling would result in more comprehensive summaries that fully preserve the plot dynamics by filling the gap between successive narrative episodes.

Integration of textual features

Another way to improve plot modeling would be to use, in addition to SNA-based representations, textual features. Basic sentiment analysis of textual content is for instance used in (Nalisnick et Baird, 2013) to polarize every relationship between characters of Shakespeare's plays, and to monitor over time the evolution of the polarity of the relation between the major characters. Similarly, (Min et Park, 2016) make use, in addition to sentiment analysis, to topic modeling based on lexical features to investigate the plot of Victor Hugo's *Les Misérables*, and to capture "topic transfers" within the plot. Though the lexical content is much sparser in movie scripts than in self-sufficient literary works, it could nonetheless provide us with additional information, even from the SNA perspective. The experiments we made with textual features when attempting to detect scene boundaries showed that named entities were well-represented in the scripts: as expected, most of these denote the characters involved in dialogue scenes, but some others denote characters or places that speakers are referring to. Such additional information could be inserted in the social network as special nodes and links in order to enhance SNA-based plot modeling.

6.2.3 Summaries

Generalization to other TV serials

As reported in Subsection 5.5.2, the evaluation of our summarization framework in a real world scenario made us focus on a single TV series. In future work, we would like to evaluate the summaries resulting from SNA-based decomposition into narrative episodes for other TV serials belonging to various genres. Furthermore, possible generalization to other kinds of video streams could be investigated, whenever the segmentation of the stream into homogeneous units can be based on alternations between different groups of interacting individuals.

Global summaries

Even though character-oriented summaries can effectively fill each user's specific information needs and benefit from the empathetic relationship viewers are likely to have with some characters, a single, global summary focusing on major characters can also be generated: most of the handmade summaries of TV serials focus simultaneously on every storyline. The extraction of the successive social neighborhoods of one character could generalize easily to sets of characters. We are then left with showing each character's typical social environment in every narrative episode, but with the objective of minimizing redundancy whenever some of the targeted characters interact with each other: each one is then part of the others' neighborhood, and extracting a few non-redundant sequences should be enough to cover such narrative episodes. Nonetheless,

global summaries must face additional issues when the plot contains multiple storylines: when editing the summary, a trade-off between chronological and narrative ordering must be found. Summarizing sequentially parallel sub-stories could result in high redundancy; conversely, introducing the selected sequences in chronological order, as they appear in the original source, may be highly confusing.

Summaries for incoming episodes

In this work, we focused on summaries of TV serials that aim at covering the past seasons to introduce the new one, but independently of its content. In contrast, summaries of the current season's past episodes are commonly found at the very beginning of each episode. Sometimes denoted as "The Previously" from its typical opening formula "previously on...", such a recap, as introducing the next episode, does not focus on the whole past plot. Instead, only some of the storylines are represented, depending on the content of the incoming episode. Careful attention to such recaps can even make the viewers figure the main narrative lines that will be developed in the next episode. For instance, the presence in such summaries of a secondary character, highly unlikely in a recap that would summarize the plot in itself, makes us guess that he is to play a role, possibly important. The generation of such a summary is then be guided by the episode to come, which acts as a filter for plot coverage. SNA-based approaches could be well-suited to generating such preparatory summaries, with a joint objective of covering past narrative sequences, but depending on the storylines represented in the next episode.

Appendix A

User study

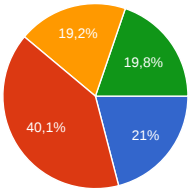
187 réponses

Afficher toutes les réponses

Résumé

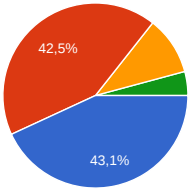
Pratiques culturelles générales

Au cours des trois derniers mois, combien de livre(s) avez-vous lus (en incluant BDs et mangas)



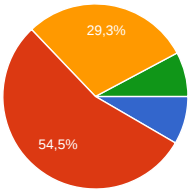
Aucun	**35**	21 %
Moins de 5	**67**	40.1 %
Entre 5 et 10	**32**	19.2 %
Plus de 10	**33**	19.8 %

Au cours des 3 derniers mois, à combien d'événements se rattachant au spectacle vivant avez-vous assisté (danse, théâtre, concert) ?



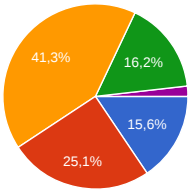
Aucun	**72**	43.1 %
Moins de 5	**71**	42.5 %
Entre 5 et 10	**17**	10.2 %
Plus de 10	**7**	4.2 %

Au cours des 3 derniers mois, combien de fois êtes-vous allé au cinéma ?



Aucune	**14**	8.4 %
Moins de 5 fois	**91**	54.5 %
Entre 5 et 10 fois	**49**	29.3 %
Plus de 10 fois	**13**	7.8 %

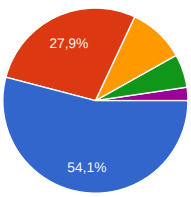
A quelle fréquence visionnez-vous des films en général ?



Tous les jours ou presque	**26**	15.6 %
3 ou 4 fois par semaine	**42**	25.1 %
1 ou 2 fois par semaine	**69**	41.3 %
1 à 3 fois par mois	**27**	16.2 %
Plus rarement	**3**	1.8 %
Jamais	**0**	0 %

Visionnage des séries en général

Au cours des 12 derniers mois, avez-vous regardé des séries TV



Tous les jours ou presque ?	**93**	54.1 %
1 à 2 fois par semaine ?	**48**	27.9 %
1 à 3 fois par mois ?	**17**	9.9 %
Plus rarement ?	**10**	5.8 %
Jamais ?	**4**	2.3 %

Quelle est la dernière série dont vous avez visionné un épisode ?

The walking dead

the walking dead

House of Cards
Gossip Girl
Daredevil
The Walking Dead
Peaky Blinders
Breaking Bad
Empire
Homeland
Devious Maids
The 100
Narcos
Grey's anatomy
House of cards
doctor who
Hannibal
Game of Thrones
The Flash
Gotham
The Wire
Scorpion
vynil
Vinyl
The Walking Dead
Death Note
Fear the Walking Dead
baron noir
Call the midwife
the 100
devious maid
black sails
the walking dead
quantico
American horror story
breaking bad
Esprits criminels
Arabesque
house of cards
The Good Wife
Supergirl
Friends
Girls
New girl
Mr Robot
deutschland 83
Farscape
New Girl
supernatural
Teen wolf
Merlin
Constantine
Glee
Blindspot
Castle
Dardevil Marvel
Dexter
Lie to me
Devious Maid
The Walking dead
The originals
Shadowhunters

How to get away with murder
iZombie
American Horror Story
The Big Bang Theory
The flash
Walking dead
Grey's Anatomy
New York section criminelle
marvel's agents of s h i e l d
Esprit criminel
desperate housewives
esprit criminel
The Big Bang Theory
Greys anatomy
Thé walking dead
Weeds
Flash
utopia
mentaliste
NCIS
Better call Saul
Elemenary
Shameless
My Mad Fat Diary
Dr Who
X-Files s01 (old school mon gars)
Broadchurch
Broken Blade
Cowboy Bebop
Kaamelott
sevda kara
Viking
better call saul
Robin des bois
Ainsi soient-ils
South Park
Scorpion sur m6
DareDevil
Arrow
Grimm
Berserk l'anime ,l'intégrale
black sails
Games Of Thrones saison 5
Prison break
Baccano
Pretty Little Liars
Devious maid
sherlock holmes
Baron Noir
Dimension W
les experts
attack of titans
Banshee
elementary
walking dead
Flash, Lucifer,
Ghost in the shell
Revenge
mad men
The big bang Theory

star wars
Shamelessí
Better Call Saul
House
Mr. Robot

Quelle est la dernière série que vous avez recommandée ?

The 100
Mr Robot
Orange is the new black
the 100
Homeland
Peaky Blinders
the walking dead
Breaking Bad
Vikings
Empire
Scandal
Rick & Morty
breaking bad
GOT
Baron Noir
game of thrones
How I met your mother
Blacklist
American Horror Story
Scorpion
Supernatural
Narcos
The Wire
Mr. Robot
Sense8
vynil
Peaky Blinders et This is England 86, 88, 90
Vinyl
narcos
True detective
scandal
peaky blinders
Twin Peaks
Gotham
The Office
X files
downton abbey
The Affair
The Flash
Orange is the New Black
Masters of Sex
Deutschland 83
OITNB + Sense 8
Ptit Quinquin
deutschland 83
Les Marvel : Agents of Shield, Jessica Jones, Daredevil
House Of Cards
DC Legends of Tomorrow
le dôme
doctor who
Hannibal
Arrow
Games Of Thrones

How to get away from murder
Dexter
Bates motel
Big Bang Theory
Devious Maid
The Shannara Chronciles
The Walking Dead
Boston Justice
Sons of anarchy
Supergirl
Quantico
Home land
The Night Manager
2 broke girls
I how met your mother
kaboul kitchen
True Detective
Murder
Game of throne
Z Nations
utopia
Pretty Little Liars
Lucifer
Borgen
Parks and recreation
Limitless
Les Chroniques de Shannara
Broken Blade
Cowboy Bebop
Peaky blinders
Viking
better call saul
24h chronos
Person of Interest
House of cards
The walking dead
Teen wolf
Erased
Games of throne
Brooklyn Nine Nine
Misfits
je ne regarde pas de série ,seulement des animes
black said
Baccano
Revenge
Devious maid
Breaking bad
sherlock holmes
attack of titans
Banshee
Jessica Jones
leftovers
walking dead
Black Mirror
Lucifer, Mr Robot
one punch man
Orphan Black
vickings
viking
the revenants

Shameless
sens 8
House of Cards
The Carmichael Show
The man in the high castle
Sherlock

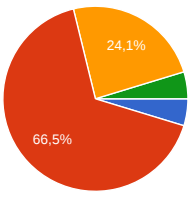
Quelles sont vos trois séries préférées ?

Kaamelott, Game of Thrones, The Wire
breaking bad, vynil
This is England 86, 88, 89 - Peaky Blinders - Family guy (les Griffin)
Empire
Je regarde peu de séries
The Walking Dead, Dexter, Grey's Anatomy
Game of Thrones, Kaamelott, Doctor Who
Scandal, Sex and the City, New Girl
Game of Throne, Breaking Bad, Fringe
game of thrones malcolm narcos
Orange is the new Black, Gossip Girl, Scandal
Breaking Bad, Homeland, Game of Thrones
desparate housewives, devious maid, malcolm
prison break, the 100, the walking dead
the 100 , the walking dead, black sails
Peaky Blinders, Secret and Lies, Once Upon a Time
peaky blinders, rectify, breaking bad
breaking bad / H / the walking dead
game of thrones spartacus quantico
Twin Peaks, Better Call Saul, Breaking Bad
Xéna, la guerrière ; Agent of shield ; Banshee
The office True blood game of thrones
breaking bad, Lost, les Simpsons
GOT ; Esprits criminels, Elementary
X files Arabesque Twin peaks
Charmed, Friends, Desperate Housewives
dowton abbey, the big bang theory, scandal
Grey's anatomy, Games of Thrones, The Good Wife
The Flash, Supergirl, The Pretty Little Liars
Friends, Charmed, Breaking Bad
Game of Thrones, New Girl, Masters of Sex
Game of thrones, Skins, How I met you mother
Grey's Anatomy, How to get away with murder, Orange is the new black
House of Cards, Girls, Mr Robot
Sense 8 - Mr Robot - Oitnb
Girls, Community, Modern Family
Agents of shield, trueblood, deutschland 83
house of cards, Downton Abbey, broadchurch
Carnivàle - Twin Peaks - Angel
Teen Wolf, The 100, Quantico
Battlestar galactica, Game of Thrones, doctor who
The 100, Orange is the new black et Gossip girl
The 100, teen wolf, the walkin dead
Desperate Housewives - Disparue - Les Revenants
Castle, Orange Is The New Black, The 100
Fringe, New Girl, Daredevil
The big bang theory. The walking dead. Flash.
Blacklist, How i met your mother, Game of thrones
Orange is the new black, GOT et Braking Bad
Les frères Scott Switched at birth Teen wolf
Friends, How I Met Your Mother et Glee
Games of Throne, Breaking Bad, House of cards

Doctor Who, Stargate, Sherlock
doctor who, the following, shaman king
Arrow, Defiance, Breaking Bad
GoT,Blacklist,Elementary
Pretty little liars, the walking dead, the 100
Stargate SG1; Games Of Thrones ; House Of Cards
Top of The Lake, Sons of Anarchy, Misfits
How to get away from murder, esprit criminel , american horror story
Walking dead, Dexter et Breaking bad
Game of Thrones, Mr Robot, Friends
Empire - orange is the new black - peaky blinders
Game of Thrones, Bates Motel, The Walking dead
Teen Wolf, sense8 et Shadowhunters
Gossip Girl Les frères scott Teen Wolf
Game of Thrones, Teen Wolf, The Shannara Chronicles
the Walking Dead, Game of Thrones, Bones
Boston Justice, The big bang theory, H
Quantico, Reign, Blindspot
Breaking bad, Orange is The New black, misfits
Dexter, Pretty Little Liars, Shameless
The Big Bang Theory, Chuck, Psych
Star Trek, Futurama, Community
OTH - Arrow - HTGAWM
Walking dead, Vampire diaries, Scorpion
Grey's Anatomy, How to Get Away with Murder, The Royals
Supernatural, Esprits criminels, New York criminelle
Games of trone; Viking; Home land
How to get away with murder, Private Practice, Dr House
Friends - Grey's Anatomy - Desperate Housewives
Desperate Housewives, Grey's anatomy, How I met your mother
Malcolm scrubs réal human
True Detective, Luther, Breaking Bad
Desperate Housewives, Big Bang Theory, Breaking bad
Arrow Flash et Gotham
The Walking Dead , Narcos , Breaking Bad
Game on throne, the walking dead, Arrow
Mr Robot, Gotham, GoT
Breaking bad, GoT, Chuck
Weeds, Games of Thrones, Breaking Bad
The Flash Arrow The Walking Dead
Scrubs , Black mirror, Community
Pretty Little Liars, Orphan Black, The Walking Dead
Heroes, Game of Thrones, Vikings
LOST, Breaking Bad, Sherlock
Breaking bad, sons of anarchy, doctor who
Weeds, Vikings, Marco Polo
The Big Bang Theory, Archer, Gotham
Borgen une femme au pouvoir ; Twin Peaks ; Ainsi soient-ils
Doctor Who, House of Cards, Friends
Community - Limitless - The Big Bang Theory
Game of Throne, Shameless , Ash vs Evil Dead
Supernatural, The 100
The Simpsons, Malcom, Futurama
Breaking Bad, Cowboy Bebop, The Wire
Game of thrones, friends, vikings
The Walking Dead, Breaking Bad et Game of Thrones
Broken Blade Vikings Psycho-Pass et Dexter ex-aequo
Cowboy Bebop, Doctor Who, Sherlock (BBC)
Vikings, Lie to me, the simpsons
OZ, The wire, Prison Break, Narcos, The Shield (oui j'en ai mis 5 mais j'ai fais un effort)

Games of thrones , The walking Dead , grey's anatomy
Viking, game of thrones, les anges
avatar le dernier maitre de l'air , breaking bad , community
Prison Break, 24h Chronos, House of Cards
Game of throne - Elementary - Black List
Friends, House MD, The Wire
Breaking Bad, The Walking Dead, South Park
J'en ai pas, je préfère les animés (°v°)
1- Game of trones 2- Reign 3- The walking dead
Sons of Anarchy-Breaking Bad-Community
Charlotte, Erased, The walking dead
Games of throne, erased,
Game of thrones, Breaking Bad, Buffy contre les vampires
Game of Thrones, The Wire, Rome
Terra nova Misfits Game of thrones
Game of thrones; Breaking Bad; Grimm
...
Rome , games of trônes , Borgia
GOT, Breaking bad , Vikings
The 100, The Vampire Diaries, Scorpion
GoT, Narcos, Brbad
Revenge, Pretty Little Liars, Once Upon À Time
Pretty Little liars, Game Of Thrones, Doctor House
game of thrones, breaking bad, sherlock holmes
Game Of Thrones, Sense8 et The Walking Dead
Friends, Homeland, Baron Noir
Orange is the new black, Jessica Jones, Game of thrones
Monogatari Series, the Walking Dead, Games of thrones
game of thrones, sherlock holmes, les experts
Gossip Girl How I met your mother 90210 Berverly Hills: nouvelle génération
walking dead , game of throne , supernatural
Game of thrones, Banshee, Person of Interest
The Walking Dead, Game of Thrones, Daredevil
lost, leftovers, six feet under
walking dead, the 100, once upon a time
The Shield, Fargo, True detective
Lucifer, Mr Robot
The walking dead, game of trône, Mr robot
Battlestar galactica, Orphan Black, Scorpion FTW
Sharpe, Downton abbey, MI-5
scrubs, boardwalk empire, game of thrones
Esprit criminel, Glee, The big bang Theory
game of thrones ; viking ; breaking bad
the walking dead, orphan black, orange is the new black (choix difficiles)
Matrioshki, niptuck, american horror story
game of throne, the big bang theory, breaking bad
The Walking Dead, Dexter, Futurama
The Wire, Breaking Bad, Louie
Game of Thrones, The 100, The Black list
suits, how I met your mother, House of cards
Sherlock, Misfits, American Horror Story
Breaking Bad, Game of Thrones et Mr. Robot

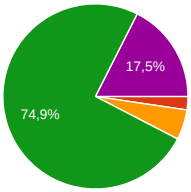
Quel genre de série préférez-vous ?



Classique : série avec une intrigue par épisode et des personnages récurrents (type Les Experts...)	**8**	4.7 %
Feuilleton : série avec une intrigue continue sur plusieurs saisons (type Game of Thrones, Breaking Bad…)	**113**	66.5 %
Mixte Classique/Feuilleton, avec une intrigue principale par épisode et une intrigue secondaire continue (type Person of Interest…)	**41**	24.1 %
Anthologie : série avec une intrigue par saison mais sans personnages récurrents d'une saison à l'autre (type American Horror Story, True Detective…)	**8**	4.7 %

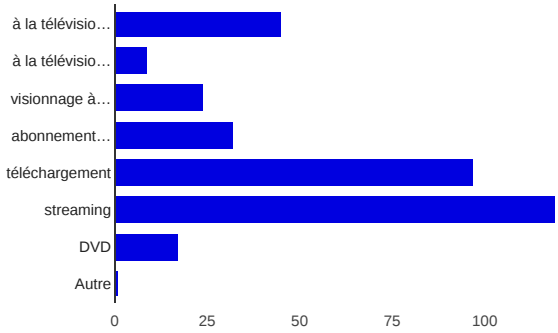
Conditions de visionnage des séries

A quel moment de la journée visionnez-vous le plus des séries ?



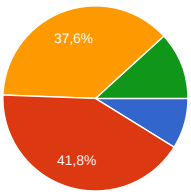
Matin	**0**	0 %
Midi	**4**	2.3 %
Après-midi	**9**	5.3 %
Soir	**128**	74.9 %
Nuit	**30**	17.5 %

Comment visionnez-vous une série le plus souvent ?



à la télévision (chaînes gratuites)	**45**	26.2 %
à la télévision (chaînes payantes)	**9**	5.2 %
visionnage à la télévision en différé (après enregistrement, en replay, …)	**24**	14 %
abonnement (Netflix, …)	**32**	18.6 %
téléchargement	**97**	56.4 %
streaming	**119**	69.2 %
DVD	**17**	9.9 %
Autre	**1**	0.6 %

Quand vous ne visionnez pas une série à la télévision en direct, en combien de temps en moyenne regardez-vous une saison ?



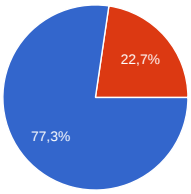
1 à 2 jours	**15**	8.8 %
1 semaine ou moins	**71**	41.8 %
2 ou 3 semaines	**64**	37.6 %
1 mois ou plus	**20**	11.8 %

Combien d'épisodes en moyenne visionnez-vous à la suite ?

3
2
4
1
5
2 ou 3
2-3
3 ou 4
6
3-4

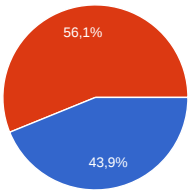
10
4-5
de 2 à 5 épisodes à la suite
1 à 2, pas plus.
8-9
dépend des sorties du jour
en général 3
4 ou 5
Deux ou trois
3 ou 4 suivant mes occupations
2 épisodes
2/3
2 ou 3 ca depend de ma fatigue
de 2 à 4
7
8
1,5
3-4 même 5
4/5
tant que je ne suis pas fatigué
Autant que possible jusqu'à qu'ils soit vraiment tard pour plus continuer
5 minimum
2-3

Visionnez-vous plusieurs séries en parallèle ?



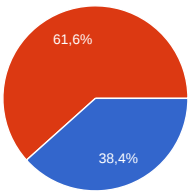
Oui	**133**	77.3 %
Non	**39**	22.7 %

Visionnez-vous des séries en faisant autre chose ?



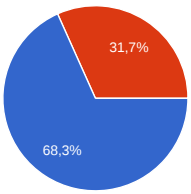
Oui	**75**	43.9 %
Non	**96**	56.1 %

Visionnez-vous certains épisodes plusieurs fois ?



Oui	**66**	38.4 %
Non	**106**	61.6 %

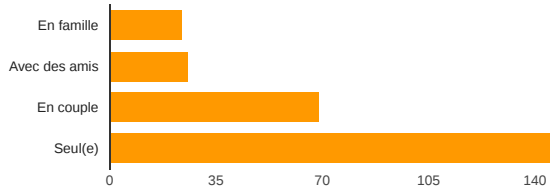
Si oui, les regardez-vous :



En intégralité	**56**	68.3 %
Par extraits	**26**	31.7 %

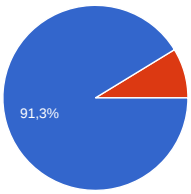
Séries et sociabilité

Avec qui regardez-vous des séries le plus souvent ?



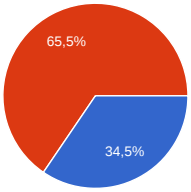
En famille	**24**	14.1 %
Avec des amis	**26**	15.3 %
En couple	**69**	40.6 %
Seul(e)	**145**	85.3 %

Parlez-vous de séries avec votre entourage ?



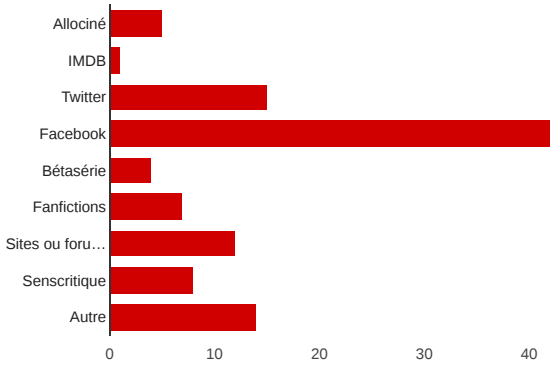
Oui	**157**	91.3 %
Non	**15**	8.7 %

Vous êtes-vous déjà exprimé au sujet de séries sur internet ?



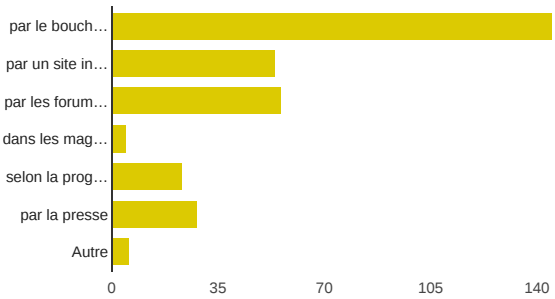
Oui	**59**	34.5 %
Non	**112**	65.5 %

Si oui, sur quelle(s) plateforme(s) ?



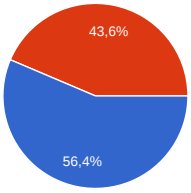
Alloci…	**5**	8.2 %
IMDB	**1**	1.6 %
Twitter	**15**	24.6 %
Facebook	**42**	68.9 %
Bétasérie	**4**	6.6 %
Fanfictions	**7**	11.5 %
Sites ou forums officiels de séries	**12**	19.7 %
Senscritique	**8**	13.1 %
Autre	**14**	23 %

Comment choisissez-vous vos séries ?



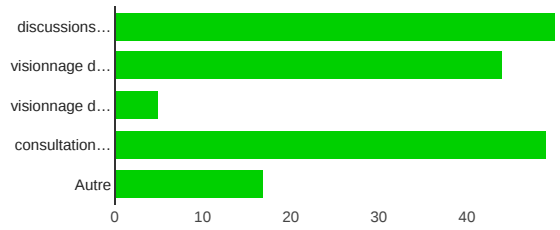
par le bouche-à-oreille	**145**	84.8 %
par un site internet dédié aux séries	**54**	31.6 %
par les forums / réseaux sociaux	**56**	32.7 %
dans les magasins culturels	**5**	2.9 %
selon la programmation TV	**23**	13.5 %
par la presse	**28**	16.4 %
Autre	**6**	3.5 %

Avant le visionnage d'une nouvelle saison de série TV avec une intrigue continue, éprouvez-vous le besoin de vous remémorer les principaux élements de l'intrigue des saisons précédentes ?



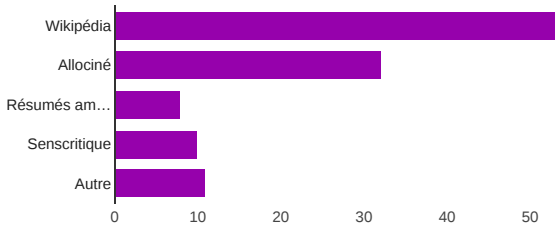
Oui	**97**	56.4 %
Non	**75**	43.6 %

Si oui, comment procédez-vous pour vous remémorer les principaux éléments de l'intrigue des saisons précédentes ?



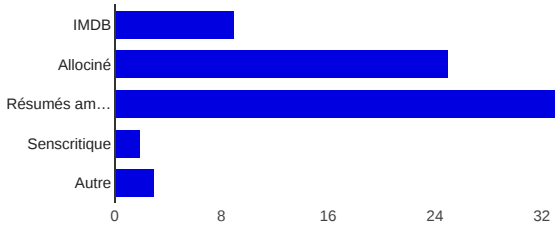
discussions avec l'entourage	**50**	48.5 %
visionnage de résumés vidéos avec extraits	**44**	42.7 %
visionnage de résumés vidéos avec extraits commentés	**5**	4.9 %
consultation de résumés texte	**49**	47.6 %
Autre	**17**	16.5 %

Si vous utilisez les résumés texte, quelle(s) plateforme(s) utilisez-vous ?



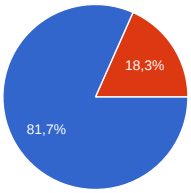
Wikipédia	**53**	63.9 %
Allociné	**32**	38.6 %
Résumés amateurs Youtube	**8**	9.6 %
Senscritique	**10**	12 %
Autre	**11**	13.3 %

Si vous utilisez les résumés vidéos sur internet, quelle(s) plateforme(s) utilisez-vous ?



IMDB	**9**	16.4 %
Allociné	**25**	45.5 %
Résumés amateurs Youtube	**33**	60 %
Senscritique	**2**	3.6 %
Autre	**3**	5.5 %

Jugez-vous utiles les résumés vidéo parfois introduits au début des épisodes de séries TV ?



Oui	**138**	81.7 %
Non	**31**	18.3 %

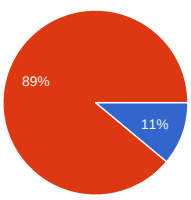
Si oui, pourquoi ?

C'est bien pour la diffusion étalée sur plusieurs semaines.
ils permettent de se remémorer les moments clés de l'épisode précédent, qui influent sur l'intrigue des autres épisodes
ça rappelle l'autre épisode
Permet de se remémorer l'intrigue
Pour se remémorer l'intrigue
Pour se remémorer les éléments des épisodes précédents si série trop complexe
Comme je ne regarde pas les résumés, celà me remet l'histoire en tête
Ils permettent de reprendre où l'on était arrêté et concentrent notre attention sur les éléments principaux de l'intrigue
On oublie des choses parfois, surtout quand il y a une longue pause entre les saisons ou les épisodes
pour se rafraichir la mémoire
utiles lorsque cela fait longtemps qu'on a visionné le dernier épisode
Ca permet de se remémorer lesz épisodes précédents
car cela nous rapelle l'épisode précédent sans devoir aller sur internet etc
Car il permet un court rappel de l'épisode précédent
Parfois, elle sont utile mais souvent elle donne une fausse interprétation
pour se remémorer les évènements précédents
Ils conviennent tout à fait à une diffusion TV (1 épisode par semaine). Les retrouver dans un coffret dvd vaut peut être moins le coup.
car ils permettent de se replonger au coeur de l'intrigue
Sauter des épisodes
Permettent de resituer l'action et de rapeller les éléments précédents (cela nous remet dans le bain)

Pour se rappeler des épisodes précédents et vérifier que l'on n'a pas déjà vu l'épisode en question
après le hiatus de plusieurs semaines, on oublie souvent des détails, donc surtout en début de saison ça sert à se rappeler les personnages etc.
Ils se focalisent sur les éléments vraiment essentiels pour comprendre l'épisode qui va suivre.
Si on ne regarde pas tout de suite l'épisode suivant ça permet de se remémorer les éléments principaux.
Parce qu'il peut parfois s'écouler beaucoup de temps avant que je trouve le temps de regarder l'épisode suivant donc ce résumé est utile pour que je me "remette dans le bain".
Si l'on ne regarde pas les épisodes de manière enchainée, cela permet de se souvenir de ce qu'il s'est passé auparavant et de ne pas se sentir perdu en reprenant la série. Cela est d'autant plus appréciable lorsqu'une nouvelle saison débute.
Quand on reprend une série et qu'on a oublié ce qu'il se passait c'est plutôt pratique !
Pour les personnes qui regardent les séries de manière lente, ou qui en visionnent plusieurs à la fois
Surtout pour le premier épisode de chaque nouvelle saison. On suit tellement de séries.
après les "season final" et les "final season", sinon d'une semaine à l'autre c'est parfois inutile
Aide-mémoire + tri infos utiles
Ça me rappelle les précédentes intrigues
Rappel des dernieres intrigues
C'est un peu rébarbatif lorsque l'on vient juste de regarder l'épisode précédent en question mais dans le cas contraire, cela permet de se recontextualiser
Pour se replonger dans l'intrigue.
Les épisodes sont souvent espacés d'une semaine et cela nous permet alors de nous remémorer des éléments et surtout d'entrer rapidement dans la mentalité de la série.
Pour se remmémorer l'intrigue et les moments clefs
Ils remettent dans l'ambiance et remémorent bien l'intrigue
Il peut se passer des mois de la sortie d'un épisode à l'autre, alors faut bien remettre les choses en place. Ou alors, il se peut que des événements bien antérieurs de la série surgissent plusieurs épisodes après.
Remet en tête les différents éléments, surtout ceux qui n'étaient pas forcément présents dans l'épisode précédent
Il n'en faut pas tout le temps, mais lorsqu'il y a eu une longue coupure entre les épisodes, parfois on oublie certains éléments
ça nous met dans le bain, ça nous replonge dans l'histoire et le contexte
Rappel rapide
rappel des choses parfois oublies
Pour remémorer ce qu'il y a eu avant
Le résumé introduit ce qu'il va se passer dans l'épisode actuel !
En cas de longue période entre deux épisodes pour tout se remémorer
Quand on espace longtemps le visionnage de deux épisodes, cela permet de se remémorer les scènes chocs des épisodes précédents
Ca nous remets dans le contexte quand nous regardons plusieurs series a la fois
Pour ne pas perdre le fil
Permet de se replonger dans l'univers de la série
Pour se remémorer ce qu'il s'est passé la semaine précédente
pour se remémorer l'intrigue
Se remettre dans le flot de l'action
Pour se remémorer, avoir une idée des éléments qui interviendront dans l'épisode et mieux comprendre la suite
Récapitulatif rapide de la situation
Quand on est sur plusieurs séries en même temps ça permet de ne pas s'embrouiller
Ils permettent une remise en contexte
Ca aide à savoir ce qu'il s'est passé auparavant. Même si on peut se douter de ce qu'il va se passer dans l'épisode.
Remémorer l'épisode précédent
Pour se remémorer certains événements importants qu'on a pu oublié
On est parfois obligés d'attendre longtemps entre 2 épisodes alors ça permet de refixer le cadre
Cela permet de ce remettre dans l'histoire
Quand on a pas pu regarder immédiatement l'épisode ou qu'il y a eu une pause cela permet de se rappeler ce qu'il s'est passé et de savoir si on a sauté un épisode par exemple
se rappeler ce qu'il s'est passer avant; resituer l'histoire
Rappeler le contexte et les derniers événements
Ils rappellent les éléments clef
En general ce qui est introduit au debut on sait qu'on va en savoir plus sur l'histoire de ces personnages ex:les freres scott ou sous le soleil
Remémoration
Rappel
me remémorer ou on en est
Ca remet dans le contexte
Beaucoup de séries suivies en même temps
Pour remettre les choses dans leur contexte et donner un premier élan scénaristique avant de rentrer au coeur de l'histoire.
Il arrive qu'on ne souvienne pas totalement de l'intrigue, surtout si on regarde plusieurs séries en parallèle

Ils permettent de se remémorer ce qu'il s'est passé avant si l'intervalle entre deux épisodes était conséquent
Généralement ils ouvrent des pistes sur les éléments qui vont être développés ; cela crée une attente et des hypothèses de visionnage.
oui dans les cas des series complexes et hébdomadaire, inutile chez netflix
Quand l'épisode sort 2 fois par mois, c'est bien de pouvoir reprendre l'intrigue.
Se remémorer les evenement passer sans revoir un épisode
Cela nous permet de nous remémorer rapidement les événments précédents
Permet d'identifier sur quels éléments les scénaristes veulent mettre l'emphase.
lorsque ça fait longtemps que j'ai pas regardé la série en question
Pour ce remettre dans le bain dans le cas d'une diffusion hebdomadaire.
Ils permettent d'avoir un résumé des élément qui vont être abordés durant l'épisode, et rien d'autre.
Car il peut arriver d'oublier un moment important de l'épisode
Pour se remémorer les principales récentes péripéties/évolution de l'intrigue
pour les pauses d'une semaine entre episode c'est utile surtout si on regarde plusieur serie a la fois
Pour mieux se suivre le deroulement de la vidéo
Pour se remémorer l'épisode précédent, tant qu'il reste court (s'il devient long par contre c'est barbant)
Pour se rappeler l'intrigue quand le dernier épisode a été visionné il y a plusieurs jours
Rien que pour entendre le "précedemment dans..." en anglais x)
Afin de ne pas perdre le fil, se mettre à jour.
Pour remémorer le chose précédentes
Pour se replonger rapidement
Ça permet de se remettre dans le bain apres une semaine de passe
Juste parce que ca remet dans l'ambiance.
Cela insiste sur les points qui vont être développés dans l'épisode et remémore des éléments lointains
Pour rememorer les elements principaux de l'intrigue
se remémorer les événements de l'épisode précédent
remttre en memoire l'intrigue
On peut voir ou on s'en ai arrêter et comme c'est la plupart du temps il y à plusieurs mois cela nous re donne l'eau à la bouche
Pour ne pas oublier des moments importants pour la compréhension de l'épisode.
Pratique si l'on regarde à la TV
Remettre dans le bain et savoir si on a pas raté l'épisode précédent
se rappeller des evenements précédants
Car quand on a plusieur séries en cour, on est directement revenu dans l'ambiance au moment de l'épisode
Ca remet dans le context l'épisode. Quelqu'un qui ne suit pas la série peut mieux comprendre ce qu'il va regarder.
Au cas où d'un oubli
cela permet de se rapeller des principaux noeuds de l'intrigue
rappel dernier épisode
pour se remémorer
ce remettre dans la série
Pour se remettre dans le bain
pour le suspens
permet de se remettre dans la serie
Ca depend
Ils permettent un meilleur suivi de l'intrigue
Ils remettent au centre parfois des enjeux que l'on pensait secondaires.
se souvenir de certains détails importants
Rappelle des détails importants pour la compréhension de l'épisode à venir

Si vous ratez un épisode, vous arrive-t-il de visionner son résumé pour passer directement à l'épisode suivant ?

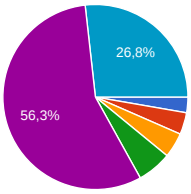


Oui	**19**	11 %
Non	**153**	89 %

Visionnage de Game of Thrones

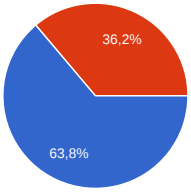
Jusqu'à quelle saison (incluse) avez-vous visionné Game of Thrones ?

Saison 1	**5**	2.7 %



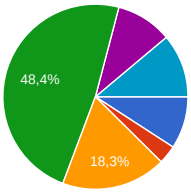
Saison 2	**7**	3.8 %
Saison 3	**8**	4.4 %
Saison 4	**11**	6 %
Saison 5	**103**	56.3 %
Aucune	**49**	26.8 %

Avez-vous visionné cette saison en totalité ?



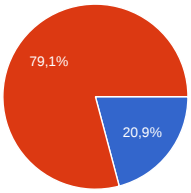
Oui	**113**	63.8 %
Non	**64**	36.2 %

Quand l'avez-vous visionnée pour la dernière fois ?



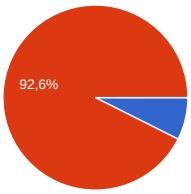
Il y a moins d'un mois	**14**	9.2 %
Il y a 2 à 3 mois	**5**	3.3 %
Il y a 3 à 6 mois	**28**	18.3 %
Il y a 6 mois à 1 an	**74**	48.4 %
Il y a plus d'un an	**15**	9.8 %
Autre	**17**	11.1 %

Avez-vous visionné une ou plusieurs saisons plusieurs fois ?



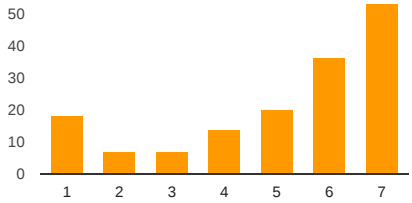
Oui	**34**	20.9 %
Non	**129**	79.1 %

Vous est-il arrivé de ne pas visionner des épisodes ou des saisons dans l'ordre ?



Oui	**12**	7.4 %
Non	**150**	92.6 %

Aimez-vous cette série ?



Pas du tout : 1	**18**	11.6 %
2	**7**	4.5 %
3	**7**	4.5 %
4	**14**	9 %
5	**20**	12.9 %
6	**36**	23.2 %
Beaucoup : 7	**53**	34.2 %

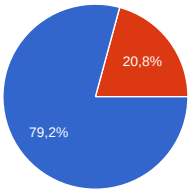
Quel est votre personnage préféré dans Game of Thrones ?

Tyrion
Tyrion Lannister
Arya Stark
Arya
Daenerys
John Snow
Aucun

tyrion
Sansa Stark
aria
Daenerys Targaryen
Aria
Jon Snow
John snow
la reine de dragon
Tirion
sans visage
la blonde avec les dragons
Daeneris Targarien / Tyrion Lannister
pas de préférence
Tyrion Lanister
Sans visage
Je ne les connais pas.
Margaery Tyrell
Snow
little finger
Jon Snow/Aria Stark
Ygritte
Tirion Lannister
Jaime
daenerys
la mechante
Jorah
Varys
Sandor Clegan (le limier)
je ne regarde pas
Aucun : je ne regarde pas !! Ça devient lourd la..!!
The Hound ! Mais comme il est mort, c'est Tormund maintenant.
C'est dur à dire ^^
Rob Stark
Eddard Stark
Ned Stark
Arya
Arya stark
Pas concerné
Jon snow
Hodor
Tyrion, Jaqen H'ghar
jon snow
Jaime Lannister
Je vous dit pas, sinon il va mourir.
Jaime et Cersei
Tyrion, of course
Tyrion lanister
Aria Stark
Arya Starc
La meuf à John snow
littlefinguer
Thyrion
Daenerys (ou aussi Margaery mais elle c'est beaucoup plus pour l'actrice que pour le personnage :)
Bronn
La fille au dragons
Aira
Oberyn Martell
Tyrion
tyrion
Turion

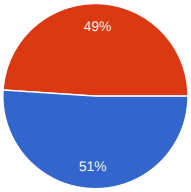
Jaqen
Therion
aucun
hodor
Eddar
snow
arya
choix trop difficile
Circee
Little Finger
Ramsay Bolton
Stanis Baratheon

Avez-vous l'intention de continuer à visionner Game of Thrones ?



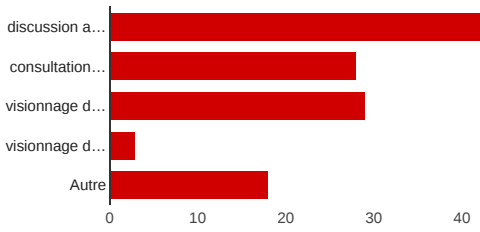
Oui	**122**	79.2 %
Non	**32**	20.8 %

Si vous pensez continuer le visionnage de Game of Thrones, éprouvez-vous le besoin de vous remémorer les principaux élements de l'intrigue des saisons précédentes ?



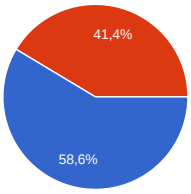
Oui	**73**	51 %
Non	**70**	49 %

Si oui, comment comptez-vous procéder pour vous remémorer les principaux éléments de l'intrigue des saisons précédentes ?



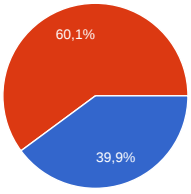
discussion avec l'entourage	**50**	58.1 %
consultation de résumés texte	**28**	32.6 %
visionnage de résumé vidéo avec extraits	**29**	33.7 %
visionnage de résumé vidéo avec extraits commentés	**3**	3.5 %
Autre	**18**	20.9 %

Connaissez-vous (approximativement) la date de sortie de la saison 6 de Game of Thrones ?



Oui	**92**	58.6 %
Non	**65**	41.4 %

Avez-vous visionné le trailer de la saison 6 de Game of Thrones ?



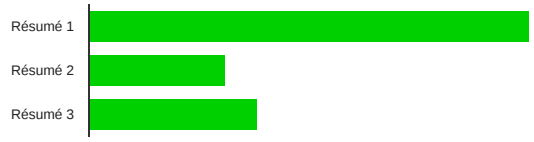
Oui	**63**	39.9 %
Non	**95**	60.1 %

Réception et évaluation de résumés automatiques pour 5 personnages de Game of Thrones

Rappelle le moins l'histoire du personnage [Pour le personnage d'Arya Stark (http://xavierbost.fr/index.php?page=arya_5), quels sont les résumés les moins satisfaisants selon les critères suivants ?]

Résumé 1	**65**	54.2 %
Résumé 2	**17**	14.2 %
Résumé 3	**38**	31.7 %

Donne le moins envie de connaître la suite de l'histoire du personnage [Pour le personnage d'Arya Stark (http://xavierbost.fr/index.php?page=arya_5), quels sont les résumés les moins satisfaisants selon les critères suivants ?]



Résumé 1	**68**	59.1 %
Résumé 2	**21**	18.3 %
Résumé 3	**26**	22.6 %

Rappelle le mieux l'histoire du personnage [Pour le personnage d'Arya Stark, quels sont les résumés les plus satisfaisants selon les critères suivants ?]

Résumé 1	**14**	12.2 %
Résumé 2	**76**	66.1 %
Résumé 3	**25**	21.7 %

Donne le plus envie de connaître la suite de l'histoire du personnage [Pour le personnage d'Arya Stark, quels sont les résumés les plus satisfaisants selon les critères suivants ?]

Résumé 1	**17**	14.7 %
Résumé 2	**66**	56.9 %
Résumé 3	**33**	28.4 %

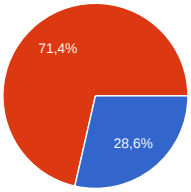
En quelques mots, quels éléments ont motivé vos choix ?

Le résumé 2 semble donner un meilleur aperçu des épisodes de la vie d'Arya Stark.
Pas du tout vu la série donc j'ai choisi celui qui me donnait une compréhensio globale
Le résumé 2 commence par l'histoire du personnage dès le départ (utile pour quelqu'un qui n'a pas vu la série)
Le résumé 2 transmet mieux le caractère du personnage alors que l'ordre des résumés 1 et 3 ne rend pas la compréhension claire
La cohérence des dialogues
le fait de découvir l'évolution du personnage
Le fait qu'on voit des éléments de toutes les saisons
Plus de descriptions, évolution physique plus visible, on la voit s'affirmer en tant que jeune femme
le résumé 1 semble trop décousu et on s'y perd
le 2 est plus complet, bonne chronologie
en comparant les différentes caractéristiques données dans chaque résumé vidéo
le changement du personnage
La chronologie des évènements et l'espacement entre ceux-ci tout au long des saisons
le fait d'intégrer les éléments de la saison 5, et de voir Arya au tout début de la série
dans le résumé 1 on ne voit pas sa vie avant la mort de son père
le n°3 s'arrête plus amplement sur les émotions d'Arya
Dans le résumé 2, on peut voir les extraits de chaque saison y compris la dernière où on l'a voit avec les sans-visages
les passages sont pas assez riches
Le personnage d'Aria est complexe et son histoire justifie plus ou moins ses actions. L'extrait 2 est celui qui m'a parut le plus exhaustif de ce point de vu là.
Le 2 est plus révélateur de son fort caractère.
caractère plus ou moins visible, plus ou moins de moments clefs constitufifs du destin du personnage
nous voyons plus d'éléments de sa vie qui semblen marquants
L'histoire d'Arya Stark est assez complexe
on voit vraiment l'évolution du personnage dans son integralité, tant au niveau de sa féminité/masculinité, de son caractère et de son courage
Principalement la chute des résumés.
Le résumé deux permet selon moi de mieux rappeler l'évolution du personnage tout en saisissant la personnalité de ce dernier.

2 plus complet
Le résumé deux qui nous souligne davantage ses origines princières.
le changement de la longueur de cheveux d'Arya Stark
le fait que le résumé 2 dresse le mieux la cntinuité et l'évolution du personnage d'Arya
Fluidité et rythme
le 3 marque l'humanité d'Arya, le 2 rappelle mieux l'histoire
j'ai pas regardé la fin (spoiler), dans le 2ème son esprit de combattante était le plus présent
Les résumés 1 et 3 omettent des événements importants et coupent les extraits trop rapidement.
chronologique
Plus d'éléments, extraits plus significatifs et qui vont plus loin dans l'histoire du personnage
Le résumé 2 suit plus l'ordre chronologique de l'histoire, ça résume bien ce qu'il s'est passé, et donne envie de voir la suite.
Le premier est plus fluide, plus centré sur la vie du personnage avec plus de détails notamment sur le début de la série on voit bien comment elle passe du statut de jeune fille à femme guerrière. On voit comment elle bascule dans la violence du monde qui l'entoure. Beaucoup plus descriptif.
La résumé 2 reprend l'introduction du personnage d'Arya dans le pilot de la série. C'est le point naturel pour commencer un résumé sur le personnage.
Resume les passages les plus emblematiques de son histoire
Les éléments choisis et leur longueur
on comprend mieux son histoire
Caractere completitude changements
Le second était une continuité qui résumait bien.
il faut montrer que Arya a soif de vengeance
les scènes montrées
Le choix des scènes montrées, les saisons montrées, le reflet de la personnalité du personnage
Les épisodes et les personnages
Que l'on voit son enfance et le fait qu'elle semble terminer au couvant
je les ai regardé rapidement , pardon..
L'initiative, les images mieux choisies pour illustrer le rôle d'Arya au sein de la série
caractère d'Arya, chronologie, personnages
Résumé 3 inutiles, Résumé 2 plus pertinent.
le début et la fin
L'initiation aux arts des sans-visage
Le résumé 2 donne plus une impression de continuité temporelle (quête initiatique du personnage), les autres semblent plus décousus.
Le résumé deux est chronologique, le résumé 3 s'arrête sur une scène qui donne envie de connaitre la suite, le résumé 1 est sans queue ni tête.
Les passages montrant sa curiosité et sa hargne
A la fois les passage choisie qui montrent plus ou moins la mentalité du personnage ainsi que le rythme et la justesse des cutes.
Le résumé 1 donne une bonne approche de l'avancement de la personnalité d'Arya et de son combat au cours des saisons, sans vraiment nous donner une envie d'en savoir plus. Le résumé 2, quant à lui, nous donne un apercu de tout ce qu'Arya a fait dans la saison, et la scène de fin du résumé donne envie de savoir ce qu'il lui arrive. Le 3° résumé est juste un mélange bizarre des deux précédents, donnant peu d'informations sur sa personne, ainsi qu'aucune envie d'en savoir plus.
Le résumé deux reprends bien l'histoire depuis le début. Le résumé 3 s'arrête en pleine action ce qui donne envie de savoir la suite.
Intégration des vidéos incompatibles avec IDM, désolé de ne pas remplir cette partie du questionnaire
Les coupures, tronçon de phrase du 1 et 2 semble plus souvent insignifiant que pour le 3. Le 3 admet plus de partie concrète. POur chacun des 3 par contre la saison 5 semble zappée (toute petite partie dans le 2).
la quantité d'informations sur le personnage que le résumé permet de remémorer
Arya et les sans-visages
La longueur des extraits et le nombre
La taille des extraits doit etre plus longue, sinon on ne comprend pas ce qu'il se passe et on ne se souvient pas
J'aime comment s'est construis sa haine pour les autres.
Le premier résumé est assez décousu, certaines scènes sont coupées
C'est le plus cohérent et il couvre toute les saison, on comprends mieux l'evolution
0
Le résumer 2 va plus loin donne des éléments plus important de l'histoire le 1 est trop rapide on à pas le temps de se rappeler à quoi correspond ce moment
Rappelle de l'histoire d'Arya sur la totalité des saisons
La psychologie du personnage, son caractere, les elements-clef de son histoire
Les points clés et passages sympathiques
va jusqu'aux evenemments recents
Plus de fluidité
Avoir un maximum d'info. Avoir des dialogues qui donnent envie de connaitre la suite. Qualité des transitions entre les scènes.
Le hasard 1d6 :D
Le 2 est plus lié, plus logique

les scènes illustrant le mieux le caractère d'Arya
le choix et la durée des scènes choisies
Un très bon début (l'introduction du personnage). Couvre plus l'histoire.
La fin du résumé 1 donne plus envie de connaitre la suite, par rapport au Limier. Le debut du résumé 2 est plus répresentatif du caractère spécial de Arya

Pour faire vos choix, avez-vous eu besoin de re-visionner certains résumés (en partie ou en totalité) ?



Oui	**34**	28.6 %
Non	**85**	71.4 %

Rappelle le moins l'histoire du personnage [Pour le personnage de Daenerys Targaryen (http://xavierbost.fr/index.php?page=daenerys_5), quels sont les résumés les moins satisfaisants selon les critères suivants ?]

Résumé 1	**41**	42.7 %
Résumé 2	**37**	38.5 %
Résumé 3	**18**	18.8 %

Donne le moins envie de connaître la suite de l'histoire du personnage [Pour le personnage de Daenerys Targaryen (http://xavierbost.fr/index.php?page=daenerys_5), quels sont les résumés les moins satisfaisants selon les critères suivants ?]

Résumé 1	**45**	48.4 %
Résumé 2	**33**	35.5 %
Résumé 3	**15**	16.1 %

Rappelle le mieux l'histoire du personnage [Pour le personnage de Daenerys Targaryen, quels sont les résumés les plus satisfaisants selon les critères suivants ?]

Résumé 1	**32**	33.3 %
Résumé 2	**30**	31.3 %
Résumé 3	**34**	35.4 %

Donne le plus envie de connaître la suite de l'histoire du personnage [Pour le personnage de Daenerys Targaryen, quels sont les résumés les plus satisfaisants selon les critères suivants ?]

Résumé 1	**17**	17.9 %
Résumé 2	**35**	36.8 %
Résumé 3	**43**	45.3 %

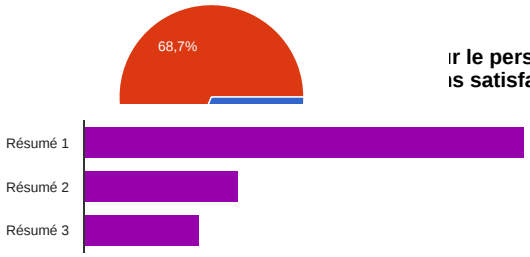
En quelques mots, quels éléments ont motivé vos choix ?

Plus clair à comprendre
Le résumé 3 montre plus d'extraits de moments importants de l'histoire de Daenerys que les autres
Compréhension des scènes
On découvre plus d'éléments sur le personnage
Le résumé 1 présente l'emsemble de l'histoire et propose plus d'éléments (on parle notamment de son armée qui est un élément clef de sa conquète de Westeros)
Le résumé 3 me semble moins bien construit, le premier me semble plus chronologique et le second résumé est mieux construit et rassemble des éléments qui attirent le spectateur dans la sphère privée du personnage
3 plus cohérente 2plus badass
en comparant les différentes caractéristiques données dans chaque résumé vidéo
apparition des personnages secondaires, le commencement du résumé
Les informations sur la vie du personnags, ses rencontres et ses pertes

plus d'informations sur la saison 5 et sur les origines du personnage
plus de détails
le 2°2 devoile la monté en puissance du personnage au fil des épisodes et saisons
La présence de Tyrion et les dragons quis ont une intrigue principale de la série
le numéro 2 montre le coté dominant et fort du personnage absent dans les autres
Les extraits plus divers montrent plus de facettes du personnage et de moments qu'il à pu vivre, cependant les extraits sont parfois mal coupés.
mis en avant des relations, presence de dragon
nous voyons + de moments différents
On ne voit aps très souvent le peronnage de Daenrys Targaryen
le fait de la voir évoluer en tant que reine, si l'on voit les dragons ou non
L'enchainement des plans parfois très saccadés.
On comprend mieux dans le résumé 3 l'évolution du personnage et sa relation avec les autres personnages. En plus les étape importantes du personnage sont présent (mort de Drago, son mariage avec une personne de l'ordre ancien...)
L'apparition tous les entourages
Car les dragons sont présents dans le résumé 2 et non dans les résumés 1 et 3. ils sont liés à l'histoire de Daenerys
Eléments de contexte et scène de fin
longueur
le troisième est le seul où on la voit bien avec son mec et ses dragons, quand même bien importants
Aucun ne m'a vraiment convaincu, les extraits sont à mon sens trop courts ou mal choisis (la scène où naissent les dragons par exemple n'apparaît dans aucun). Le 1 m'a paru mieux car extraits plus longs et mieux choisis, sans conviction
La longueur des séquences et les moments choisis
manque d'elements cles
Son histoire de reconquérir son pays
les scènes montrées
Le choix des scènes montrées, les saisons montrées, le reflet de la personnalité du personnage
Montrer les qualités et les pouvoirs de la Khalicie
dragons, caractère, personnages
Notamment les dragons
Le résumé 3 donne un bon aperçu de l'évolution du personnage au fil du temps (fragile d'abord, puis plein d'assurance plus tard). Des 3, c'est le plus intéressant, même s'il omet certains éléments factuels importants.
La représentation de son côté autoritaire mais juste
Résumé 1 : introduction au personnage réussie, mais aucune piste qui puisse nous ramener pour en découvrir plus. Résumé 2 : Du suspens, avec très peu d'histoire du personnage. Résumé 3 : bof.
Aucun n'est ici assez pertinent. Si on résume Daenerys en 3 mots : mère des dragons. C'est l'association principale que l'on apporte a Daenerys. Dans aucun des extraits on ne voit vraiment qu'elle possède des dragons (hormis quelques secondes dans le 2 mais sans trop vraiment comprendre). DOnc la principale caractéristique de Daenerys n'est vraiment présente dans aucun.
Son amour pour Drogo et sa rencontre avec Tyrion
En faite les 3 sont pas top, c'est trop court
Pas assez de details
Le début de l'histoire n'est pas assez détaillé dans le 2ème résumé
cohérent et il couvre toute les saison, on comprends mieux l'evolution
0
Les élements rappellés sont plus complets
Les resumés 1 et 3 ne parlent meme pas des dragons
Pareil que précédemment mais il manque quelques points clés comme acquisition de l'armée et des dragons
moments clefs, extraits assez longs pour se souvenir
des points importants de sa vie. Avoir des informations sur l'histoire sans spoiler.
Sacrée hasard !
Plus chronologique
le choix des scènes
le choix et la durée des scènes choisies
Resumé 1: apparition de Tyrion qui est importante, Resumé 2 : dragon
Les dragon ne sont que dans le résumé 2

Pour faire vos choix, avez-vous eu besoin de re-visionner certains résumés (en partie ou en totalité) ?

Oui	**31**	31.3 %
Non	**68**	68.7 %



ır le personnage de Jaime Lannister (http://xavierbost.fr/index.php?
ıs satisfaisants selon les critères suivants ?]

Résumé 1	**57**	62 %
Résumé 2	**20**	21.7 %
Résumé 3	**15**	16.3 %

Donne le moins envie de connaître la suite de l'histoire du personnage [Pour le personnage de Jaime Lannister (http://xavierbost.fr/index.php?page=jaime_5), quels sont les résumés les moins satisfaisants selon les critères suivants ?]

Résumé 1	**49**	53.8 %
Résumé 2	**24**	26.4 %
Résumé 3	**18**	19.8 %

Rappelle le plus l'histoire du personnage [Pour le personnage de Jaime Lannister, quels sont les résumés les plus satisfaisants selon les critères suivants ?]

Résumé 1	**18**	19.4 %
Résumé 2	**36**	38.7 %
Résumé 3	**39**	41.9 %

Donne le plus envie de connaître la suite de l'histoire du personnage [Pour le personnage de Jaime Lannister, quels sont les résumés les plus satisfaisants selon les critères suivants ?]

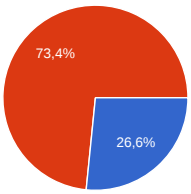
Résumé 1	**17**	18.3 %
Résumé 2	**34**	36.6 %
Résumé 3	**42**	45.2 %

En quelques mots, quels éléments ont motivé vos choix ?

Phrases plus liées
Meilleure organisation des scènes
Le plus de détail concernant le personnage (sa relation avec sa soeur et son frère) ainsi que sa fille.
le résumé 2 montre moi mais raconte plus l'histoire du personnage, le 1° donne en revanche plus envie de suivre ce personne car il y a un peu plus de suspens dans la construction
3a plus de sens
la longueur des scènes
Les différents évènements présentés et les personnages rencontrés
l'histoire est plus détaillée dans le 3, mais le 2 est bien plus intriguant
l'ensemble des details
Car c'est celui qui a l'air de respecter le plus la chronologie de la serie
Son combat contre les Stark, sa main coupée, le désir de vengeance quand sa "fille" se fait empoisonner
les passages sont plus long et mieux choisis
Dans l'extrait 3 Jaime Lannister est plus humanisé et la progression de sa personnalité est plus montrée.
La fin du deux avec l'intringue sur le sort de sa fille donne envie de connaître la suite, le 3 revient bien sur les éléments marquants son histoire.
plus d'élément, manque de visibilité de la dualité du personnage
le personnage apparaît sous des aspects différents
Au fil des saisons, ce personnage devient de plus en plus humaniste
le suspens présent, les details plus pertinants
Les éléments de l'intrigue dévoilés
Résumé 1: Ne donne pas assez d'élément marquant sur le personnage ; résumés 2 et 3 sont les plus juste sur le rappel des élément marquant. Les 3 et 2 permette de mieux replacer dans le contexte général de la série le personnage.
Les intrigues différentes
Dans le résumé 2 certains extraits sont trop courts ou coupés trop brusquement
Impression faux raccords et continuité sonore

problème de visionnage video2
Le résumé 1 se clôt avec un extrait non significatif et surtout en pleine phrase, même s'il rappelle bien l'histoire du personnage. C'est le contraire pour le résumé 2 (extraits bien trop courts et se coupent en plein dialogue). Je n'ai pas toujours aimé le choix des scènes pour le 3ème, il manquait des scènes importantes à mon sens. Encore une fois, aucun des résumés ne m'a vraiment convaincue.
Le résumé 2 est plus complet et aborde les éléments les plus importants concernant Jaime.
Le moment où Myrcella va mourir
trop plat
On comprend mieux quand on voit le début de l'intrigue
les scènes montrées
Le choix des scènes montrées, les saisons montrées, le reflet de la personnalité du personnage
personnages, chronologie
Mort de Myrcella
Le résumé 1 rappelle mieux les liens de Jaime avec les autres membres de sa famille. Le résumé 2 est le plus décousu. Le résumé 3 a une bonne dynamique quant à la chronologie, on saisit bien les différentes étapes de l'évolution du personnage.
Le résumé 2 montre la cruauté de Jaime qui est plutôt un personnage brave et juste en théorie
Beaucoup plus de moment clé dans le 2 je trouve
Le changement de caractère de Jaime
La longueur
Extrais plus long et plus comprehensibles
Il manque certaines scènes clés comme celle avec Bran au début (pas dans le 2ème résumé) ou la mort de Myrcella (pas dans le 1er résumé)
cohérent et il couvre toute les saison, on comprends mieux l'evolution
0
rappelle les elements principaux
J'hésite entre le 2 et le 3 mais pas avec le 1 car il manque la mort de myrcella
scènes completes
Que les scènes ne soient pas trop rapides, trop coupées.
...
Logique des événements
les différentes rencontre de JL
Resumé 3 contient un moment important : la rencontre avec cercei avec sa main coupée.

Pour faire vos choix, avez-vous eu besoin de re-visionner certains résumés (en partie ou en totalité) ?



Oui	**25**	26.6 %
Non	**69**	73.4 %

Rappelle le moins l'histoire du personnage [Pour le personnage de Sansa Stark (http://xavierbost.fr/index.php?page=sansa_5), quels sont les résumés les moins satisfaisants selon les critères suivants ?]

Résumé 1	**45**	48.9 %
Résumé 2	**27**	29.3 %
Résumé 3	**20**	21.7 %

Donne le moins envie de connaître la suite de l'histoire du personnage [Pour le personnage de Sansa Stark (http://xavierbost.fr/index.php?page=sansa_5), quels sont les résumés les moins satisfaisants selon les critères suivants ?]

Résumé 1	**35**	39.3 %
Résumé 2	**29**	32.6 %
Résumé 3	**25**	28.1 %

Rappelle le mieux l'histoire du personnage [Pour le personnage de Sansa Stark, quels sont les résumés les plus satisfaisants selon les critères suivants ?]

Résumé 1	**33**	36.7 %
Résumé 2	**21**	23.3 %
Résumé 3	**36**	40 %

ı personnage [Pour le personnage de Sansa Stark, quels sont les ts ?]

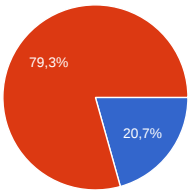
Résumé 1	**26**	28.6 %
Résumé 2	**20**	22 %
Résumé 3	**45**	49.5 %

En quelques mots, quels éléments ont motivé vos choix ?

Séquences plus explicites, moins coupées
Plus clair
Le résumé 3 est le plus détaillé
On voit plus l'évolution physique de Sansa et on apprend plus de choses sur sa vie que dans le premier résumé. Cependant, le premier résumé est bien construit et laisse justement une part de mystère qui donne envie d'en découvrir plus. Le troisième résumé est moins explicite et moins attrayant pour le téléspectateur selon moi
en me basant sur l'expérience racontée du personnage
les paroles du personnage
Les dialogues et les évènements
toujous plus d'éléments de la saison 5 dans le 3, plus de matière, mais le 2 était bien car on comprenait l'histoire de la famille de Sansa
détails
Sur le n°3 on voit plus que sur les autres Sansa se transformé entre le début et la fin du résumé
La dernière scène avec Théon est intéressante et on comprend mieux l'histoire de la famille Stark
nous voyons mieu les différentes facettes de sansa
Sansa est plus humanisée dans le résumé numéro 3.
C'est le résumé le plus court et le mieux construit avec tous les éléments caractéristiques du personnage.
naiveté, regret
Sans Stark murît de plus en plus au fil des saison, ce peut être un personnage intéressant pour la saison 6
le résumé 3 donne plus d'indications sur l'évolution de Sansa, on la voit aussi bien sensible que puissante et téméraire
Le résumé 1 et 3 explique au mieux l'évolution du personnage et surtout sont évolution dans les différentes maisons.
Les inrigues divers
Scène de début initiatique et même scène de fin
longueur, hachure, incohérence
tous bien!
Je ne suis de nouveau pas convaincue car à mon sens il aurait fallu terminer par l'évasion de Sansa avec Theon pour donner envie de regarder la suite. Mais le résumé 3 montre des scènes significatives de l'évolution du personnage tout en allant le plus loin dans l'histoire du personnage à la saison 5.
Le résumé 3 aborde plus les dernières saisons: ça donne plus envie de découvrir la suite.
Les moments choisis
difficile a dire, trop semblables
On doit voir tout son parcours
les scènes montrées
l'ordre d'apparition des scènes
Le choix des scènes montrées, les saisons montrées, le reflet de la personnalité du personnage
personnages, chronologie
Margeary, mort de son père...
Le résumé 1 développe mieux la personnalité de Sansa et met le mariage au premier plan, problématique centrale. Le résumé 2 est très décousu, même s'il met fortement l'accent sur Joffrey. Le résumé 3 s'attarde plus sur les personnages satellites (Cersei, Baelish), mais rate du même coup les intrigues relatives au mariage.
résumé 1 mieux pour les premières saisons, pour les dernière saisons le 3 est mieux je trouve
Son mariage avec Tyrion
Le nombre d'extrait
Plus de details. Mais il faut que cela soit plus long
Plus de scènes importantes dans le 3ème, surtout vers la fin
cohérent et il couvre toute les saison, on comprends mieux l'evolution
0
Les points clés
Que les paroles aient du sens entre les scènes. Informations clefs.
Faits chronologiques, plus complet

les scènes clés
Plus coherent. Donne plus d'info sur la dernière saison

Pour faire vos choix, avez-vous eu besoin de re-visionner certains résumés (en partie ou en totalité) ?



Oui	**19**	20.7 %
Non	**73**	79.3 %

Rappelle le moins l'histoire du personnage [Pour le personnage de Theon Greyjoy (http://xavierbost.fr/index.php?page=theon_5), quels sont les résumés les moins satisfaisants selon les critères suivants ?]

Résumé 1	**40**	46 %
Résumé 2	**28**	32.2 %
Résumé 3	**19**	21.8 %

Donne le moins envie de connaître la suite de l'histoire du personnage [Pour le personnage de Theon Greyjoy (http://xavierbost.fr/index.php?page=theon_5), quels sont les résumés les moins satisfaisants selon les critères suivants ?]

Résumé 1	**46**	52.3 %
Résumé 2	**27**	30.7 %
Résumé 3	**15**	17 %

Rappelle le plus l'histoire du personnage [Pour le personnage de Theon Greyjoy, quels sont les résumés les plus satisfaisants selon les critères suivants ?]

Résumé 1	**19**	20.9 %
Résumé 2	**35**	38.5 %
Résumé 3	**37**	40.7 %

Donne le plus envie de connaître la suite de l'histoire du personnage [Pour le personnage de Theon Greyjoy, quels sont les résumés les plus satisfaisants selon les critères suivants ?]

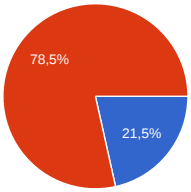
Résumé 1	**15**	16.7 %
Résumé 2	**28**	31.1 %
Résumé 3	**47**	52.2 %

En quelques mots, quels éléments ont motivé vos choix ?

Phrases moins coupées
moins sacadé
L'histoire est la plus déaillée dans le résumé 3, on voit bien sa relation avec les autres personnages et le basculement vers Schlingue
Le second résumé est dans son ensemble beaucoup mieux ficelé que les deux autres résumés.
Les évènements, leur pertinence
mêmes raisons que précédemment, ici le 3 montre une beaucoup d'éléments de toutes les saisons dont la saison 5
details
C'est dans le n°1 qu'on voit le mieux la descente aux enfers de Theon
Théon évoluent et trahis les siens. On a de suite envie de savoir pourquoi et s'il va changer une nouvelle fois
plus d'importance à son début dans le nord
La fin du 2 permet du suspens quant au sort réserver au personnage.
on ne comprend l'évolution du personnage que partiellement
nous voyons plus d'éléments de son histoire
Theon Greyjoy n'est pas un personnage très important à la série
de voir le personnage passer du pouvoir à la soumission

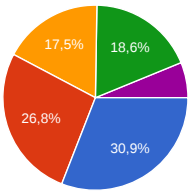
l'enchainement des révélations sur l'histoire du personnage
résumé deux : Rappel au mieux l'évolution du personnage et sa relation avec les autre personnages .
Les plus des intrigues
le résumé 3 est trop court
Eléments de contexte et scène de fin
rien ne peut donner envie de savoir plus sur ce personnage
Encore une fois, j'aurais préféré une autre scène finale. Le résumé 3 montrait le mieux, à la fois la relation de Theon avec Robb, son origine familiale, et sa transformation en "Schlingue"
Les moments choisis et l'ordre dans lequel ils sont montés
trop semblables aussi
Je trouve que les 3 extraits se ressemble mais on voit toutes les péripéties
les scènes montrées
Même réponse que pour Sansa
Le choix des scènes montrées, les saisons montrées, le reflet de la personnalité du personnage
personnages, personnalité
Le résumé 1 illustre bien la problématique de légitimité pour Theon (problématique de filiation). On voit bien la phase de déchéance. Le résumé 2 se focalise plus sur les relations avec sa soeur, et présente plus finement la problématique du pouvoir/de l'autorité. Le résumé 3 fait une transition trop brutale entre la phase ascendante de Theon et sa déchéance.
PLus de scene importante il me semble
manque d'éléments importants dans son histoire
La trahison de Théon envers Robb
Longueur
J'ai trouvé le troisième plus décousu que les autres, avec des scènes coupées
cohérent et il couvre toute les saison, on comprends mieux l'evolution
0
Les elements-clef
Dans le 3 il indique clairement les choses
un peu au feeling.
Plus complet et résumé, plus dramatique
les scènes ou Theon change
resume plus l'evolution du personnage

Pour faire vos choix, avez-vous eu besoin de re-visionner certains résumés (en partie ou en totalité) ?



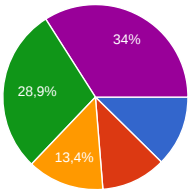
Oui	**20**	21.5 %
Non	**73**	78.5 %

Parmi les 5 personnages, quel est celui dont les résumés vous ont le plus donné envie de visionner la suite de Game of Thrones ?



Arya Stark	**30**	30.9 %
Daenerys Targaryen	**26**	26.8 %
Jaime Lannister	**17**	17.5 %
Sansa Stark	**18**	18.6 %
Theon Greyjoy	**6**	6.2 %

Parmi les 5 personnages, quel est celui dont les résumés vous ont le moins donné envie de visionner la suite de Game of Thrones ?



Arya Stark	**12**	12.4 %
Daenerys Targaryen	**11**	11.3 %
Jaime Lannister	**13**	13.4 %
Sansa Stark	**28**	28.9 %
Theon Greyjoy	**33**	34 %

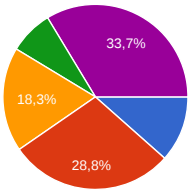
Perception de la durée

Selon vous, quelle était la durée moyenne des résumés ?

2 minutes
3 minutes
2 min
2min
1 minute 30
2 minutes 30
2mn30
deux minutes
2 min 30
2m10
3 minutes
1min20
2 minutes
3min
5 min
2minutes
2 min 40
1min45
2min20
Entre 2 minutes 30 et 3 minutes
2 mintues et demi
2,10 min
2.30 minutes
Trop long, un résumé selon moi doit être inférieur à 1min. Surtout qu'il s'agit là d'un personnage et pas d'un ensemble de protagonniste. Ou alors moins de deux minutes mais avec un rythme plus lent. Adapté à une lecture confortable.
2:26
1.50
1:30
3mn
2
2mn
25 mins
Raisonnable
1 à 2 minutes
2:30
2 à 3 minutes
4
Etant au début de la 4ème saison j'ai préféré ne pas visionner les résumé pour éviter de gâcher les deux saisons qui me restent avant la sortie de la 6ème. Désolé...
3
1 minutes 30
1:50
2:20
Oups... J'ai lu les durées par réflexe :/ Mais trop courts
1 à 3 min
De ceux d'Arya, 2m30. Trop d'extraits, pas le temps de tout voir.
7 min
1 min 30
entre 2 et 3 minutes
1min
1 minute 30 sec
2.5
0
2:30 min
5"
3 minutes 30
8
Trop long.
1'30

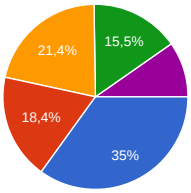
2min30

Selon vous, quel était le personnage dont les résumés étaient les plus courts ?



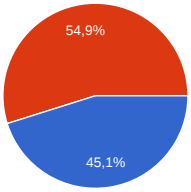
Arya Stark	**12**	11.5 %
Daenerys Targaryen	**30**	28.8 %
Jaime Lannister	**19**	18.3 %
Sansa Stark	**8**	7.7 %
Theon Greyjoy	**35**	33.7 %

Selon vous, quel était le personnage dont les résumés étaient les plus longs ?



Arya Stark	**36**	35 %
Daenerys Targaryen	**19**	18.4 %
Jaime Lannister	**22**	21.4 %
Sansa Stark	**16**	15.5 %
Theon Greyjoy	**10**	9.7 %

Possédez-vous une montre ?



Oui	**65**	45.1 %
Non	**79**	54.9 %

Fiche d'identité

Quel âge avez-vous ?

19
22
21
20
20 ans
24
23
18
25
26
19 ans
22 ans
24 ans
20 ans
35
28
27
23 ans
21 ans
29
18 ans
33
39
38
15 ans
40
23ans
44

20ans
25 ans
48
27 ans
17
22ans
34
18 ans
36
31
52
47
31 ans
32

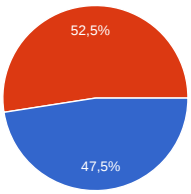
Quelle est votre formation universitaire ?

Master
Licence
M1 PCC
Informatique
Licence Informatique
Licence informatique
L3 Information et Communication
info-com
Licence Info-com
Master LEA
Histoire
L1
LLCER Anglais
Lettres
Licence Informatique
dut stid
DUT STID
Doctorat
doctorat
médiation culturelle
Master 1 Stratégie du développement culturel - Publics de la culture et de la communication
l3 info comm
M1 Publics de la Culture et Communication
Licence 3 Information - Communication
Master 1 Public de la Culture et Communication
M1: PCC
Master PCC
l3 info com
L3 Information & Communication
licence 3 information communication
L3 information & communication
information-communication
Licence 3 Information Communication
L3 information et communication
information communication
Info - Com
L3 information-communication
Information et communication
licence info comm
l3 infocom
Licence 3 Information & communication
L3 info-com
Master 1 Culture & communication
L3 Information et Communication (Avignon)

master publics de la culture et communication
Public de la culture et communication
communication
Doctorat
L3 Info-Com
histoire-géographie
Bac +5, Ecole de commerce
M1 Langues
L3 anglais
Master 2 PCC
Classe prépa + IEP
Master 2 Communication
Licence de Lettres modernes
2nde
Bac +8
licence
metiers de la culture
Licence ABF
Master LEA Action Humanitaire
L1 info-com
Commerce international
Master 2 en Economie sociale et solidaire
L3 droit
Licence InfoCom et Master Direction Artistique
Ecole d'ingénieur
Lettres modernes
Tourisme
L 1
Quel intérêt pour le questionnaire ?
ingenieur
L2LLCER ITALIEN
Premiere année de licence
Licence Arts Lettres et Langues
Bts tourisme
Sciences
L1 lea
Master 2 cultures et sociétés italiennes
DUT TC
l1 info-com
Dut
Licence Lettres-langues
Master Publics de la culture et communication
LEA Anglais- Espagnol
LEA
2ème année Licence Information-Communication
LLCER anglais
Ufr STS
licence pro en communication
Droit
licence
Master Communication d'Entreprise
master
Licence info-com
master
Doctorat (en cours)
Master ILSEN
L3 aes
L1 Info
CERI L1 Info
Info com et Histoire

M1-geographie
L2 Info
L1 Informatique
l2
L2 (en cours)
M2 Droit International, M2 Expertise pénale, L1 Informatique
Master géomatique
Master 1 ILSEN
En cours d'acquisition de Licence 1
Licence informatique première année
Bac+3
M2
Pmbo
Master 2
bac +5
Docteur en informatique
Licence info
Master Informatique
Biologie
Doctorat en biologie
CPGE
SVT
informatique
L1 CERI Avignon
Dut statistiques et informatique decisionnelle
L1 CMI BIOLOGIE
RISM M1
bac+5 socio
M1 informatique RISM
bac +5 informatique
Master 1 Informatique
l1 info
bac +2
master RISM
Master RARE
informatique
maitrise
Doctorat géoscience
license informatique
formation scientifique
Licence Marketing
Doctorant Cinéma (recherche)
Doctorant Informatique
Licence anglais
LMD

Vous êtes :



Un homme	**87**	47.5 %
Une femme	**96**	52.5 %

Quel est le code postal de votre résidence actuelle ?

84000
84140
30133
31000

84700
84100
13150
73000
30400
84270
84
35000
44000
38000
13910
34000
30150
8400
84200
84310
13990
84000
84130
30650
44300
49000
75015
75018
14000
59300
22100
31100
54500
62880
09000
13870
Quel intérêt ?
84250
84800
84300
44200
38690
30131
84470
26120
30210
84450
30290
84120
84510
BL1 4RL
37000
84170
13008
84740
26200
84210
67000
13560
13160
italie
78640
84850
77

75002
75017

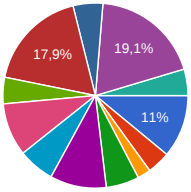
Quel est le code postal de la ville où vous avez vécu le plus longtemps ?

84000
30133
84700
13150
84210
30150
84500
04150
84100
84130
84600
84800
34000
13870
30650
84300
73000
30200
84120
84510
8400
84200
31000
69000
75015
35530
25700
84310
63112
13500
33290
13990
84270
13320
30190
30400
34120
13960
38140
07200
78660
300380
28400
13
44110
29440
14167 (All.)
17580
33190
53200
75017
53240
59300
38000
22100
50690

26170
95240
13490
44270
62880
22380
85170
Quel intérêt ?
5109
84
84250
50180
07150
26110
74600
53000
x
13800
13750
30700
13810
101 (Madagascar)
97417
38690
50360
49000
99
30900
84450
13890
84220
07350
26120
04700
13920
35290
84170
74200
77000
07700
Maroc
13008
43000
35000
30330
84740
25000
40140
99999
84110
26200
13110
Algerie 06035
75009
74000
13560
30130
13570
13200
51100

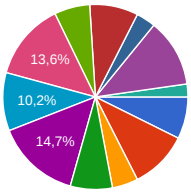
84850
94
93140
64000
14000
16000
Etranger
19100

Quel est le niveau d'étude de votre père ?



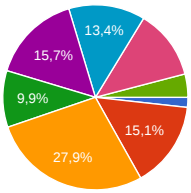
CAP	**19**	11 %
BEP	**7**	4 %
BTS	**4**	2.3 %
Bac professionnel	**10**	5.8 %
BAC	**17**	9.8 %
Bac +2	**11**	6.4 %
Bac +3	**16**	9.2 %
Bac +4	**8**	4.6 %
Bac +5	**31**	17.9 %
Doctorat	**9**	5.2 %
Sans diplôme	**33**	19.1 %
Autre	**8**	4.6 %

Quel est le niveau d'étude de votre mère ?



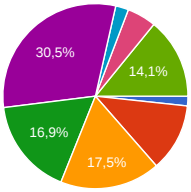
CAP	**13**	7.3 %
BEP	**18**	10.2 %
BTS	**8**	4.5 %
Bac professionnel	**13**	7.3 %
BAC	**26**	14.7 %
Bac +2	**18**	10.2 %
Bac +3	**24**	13.6 %
Bac +4	**11**	6.2 %
Bac +5	**15**	8.5 %
Doctorat	**6**	3.4 %
Sans diplôme	**21**	11.9 %
Autre	**4**	2.3 %

Quelle est la catégorie professionnelle de votre père ?



Agriculteurs exploitants	**3**	1.7 %
Artisans, commerçants et chefs d'entreprise	**26**	15.1 %
Cadres et professions intellectuelles supérieures	**48**	27.9 %
Professions Intermédiaires	**17**	9.9 %
Employés	**27**	15.7 %
Ouvriers	**23**	13.4 %
Retraités	**21**	12.2 %
Autres personnes sans activité professionnelle	**7**	4.1 %

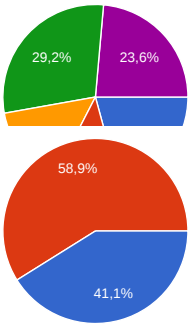
Quelle est la catégorie professionnelle de votre mère ?



Agriculteurs exploitants	**3**	1.7 %
Artisans, commerçants et chefs d'entreprise	**21**	11.9 %
Cadres et professions intellectuelles supérieures	**31**	17.5 %
Professions Intermédiaires	**30**	16.9 %
Employés	**54**	30.5 %
Ouvriers	**4**	2.3 %
Retraités	**9**	5.1 %
Autres personnes sans activité professionnelle	**25**	14.1 %

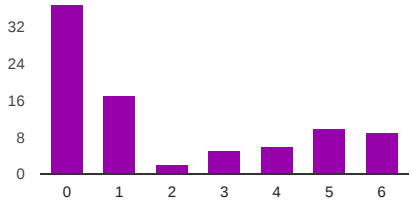
Durant l'année universitaire quelle est la situation qui se rapproche le plus de la vôtre ? Vous avez :

une activité rémunérée à plein temps	**37**	20.8 %
une activité rémunérée à mi-temps au moins 6 mois par an	**21**	11.8 %
une activité rémunérée occasionnelle	**26**	14.6 %
aucune activité rémunérée sauf l'été	**52**	29.2 %
aucune activité rémunérée	**42**	23.6 %



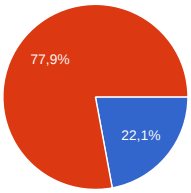
Oui	**74**	41.1 %
Non	**106**	58.9 %

Si oui, à quel échelon ?



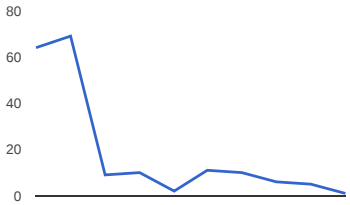
Echelon 0 : 0	**37**	43 %
1	**17**	19.8 %
2	**2**	2.3 %
3	**5**	5.8 %
4	**6**	7 %
5	**10**	11.6 %
Echelon 6 : 6	**9**	10.5 %

Où avez-vous rempli ce questionnaire ?



En salle informatique 2W13	**36**	22.1 %
Autre	**127**	77.9 %

Nombre de réponses quotidiennes



Appendix B

Cumulative networks

B.1 Breaking Bad

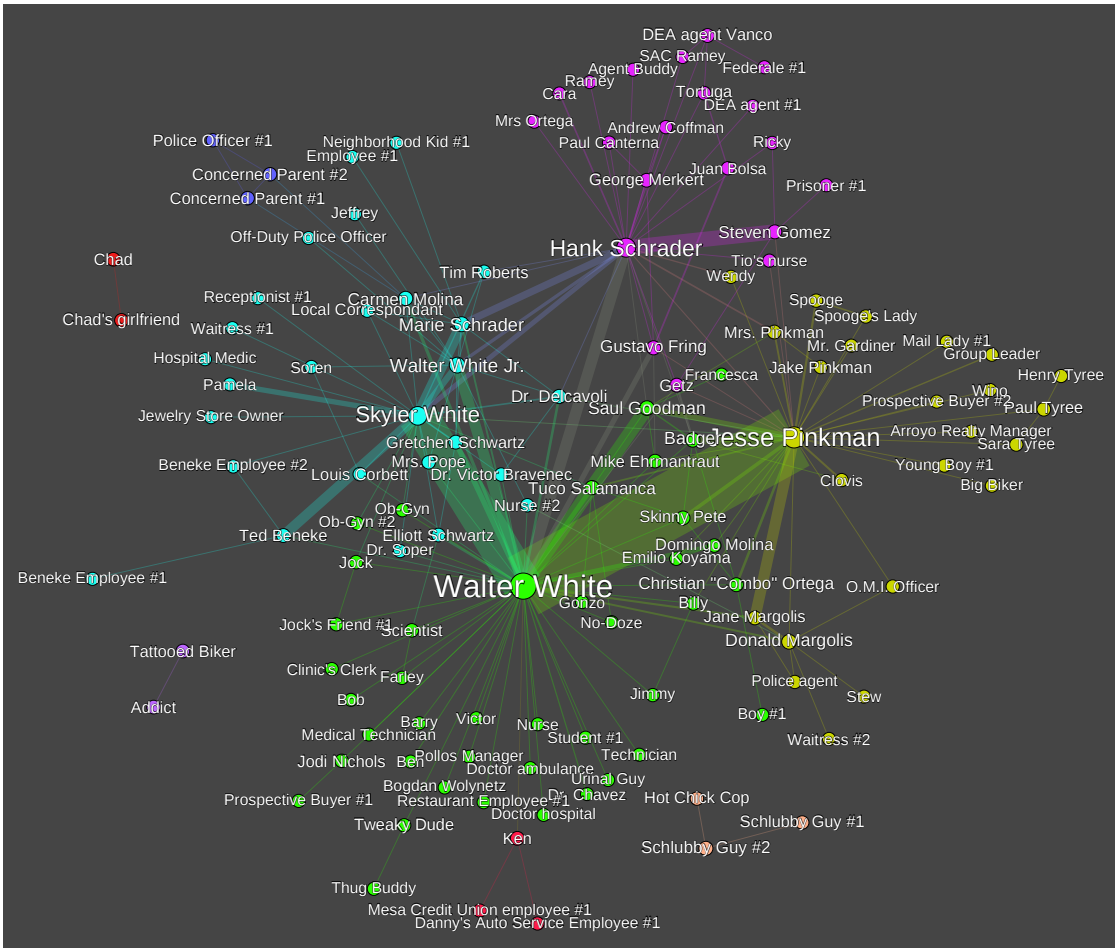


Figure B.1 – *Cumulative network for the first two seasons of* Breaking Bad.

B.2 Game of Thrones

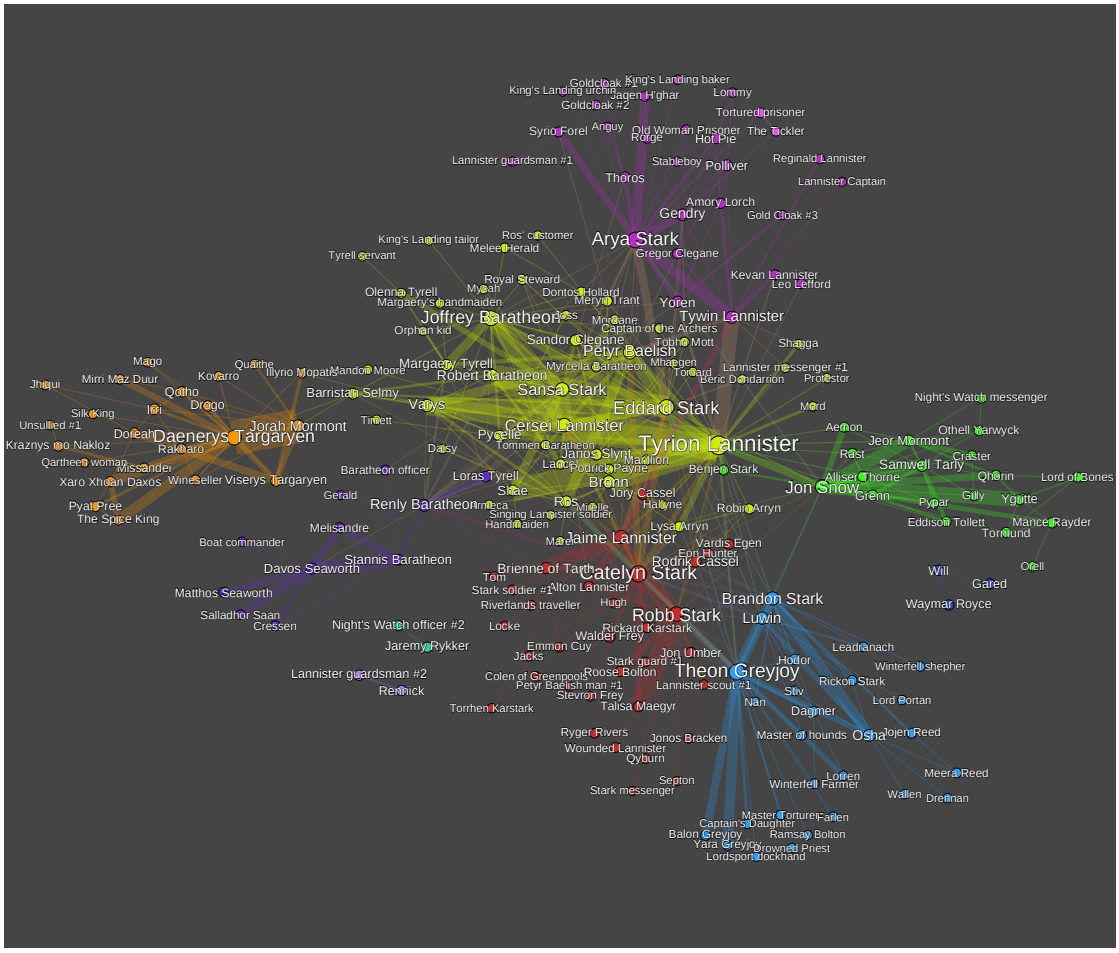


Figure B.2 – *Cumulative network for the first two seasons of* Game of Thrones.

B.3 House of Cards

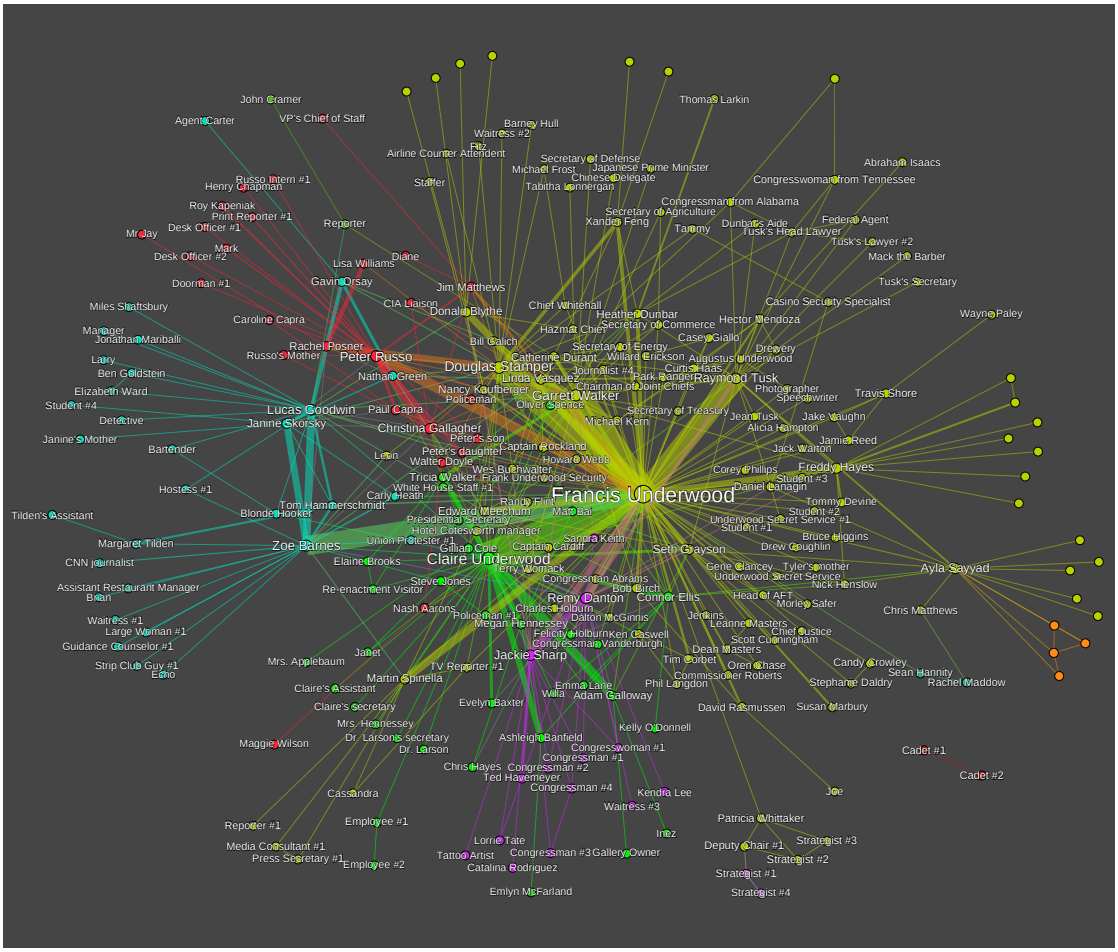


Figure B.3 – *Cumulative network for the first two seasons of* House of Cards.

Appendix C

Characters' social storylines

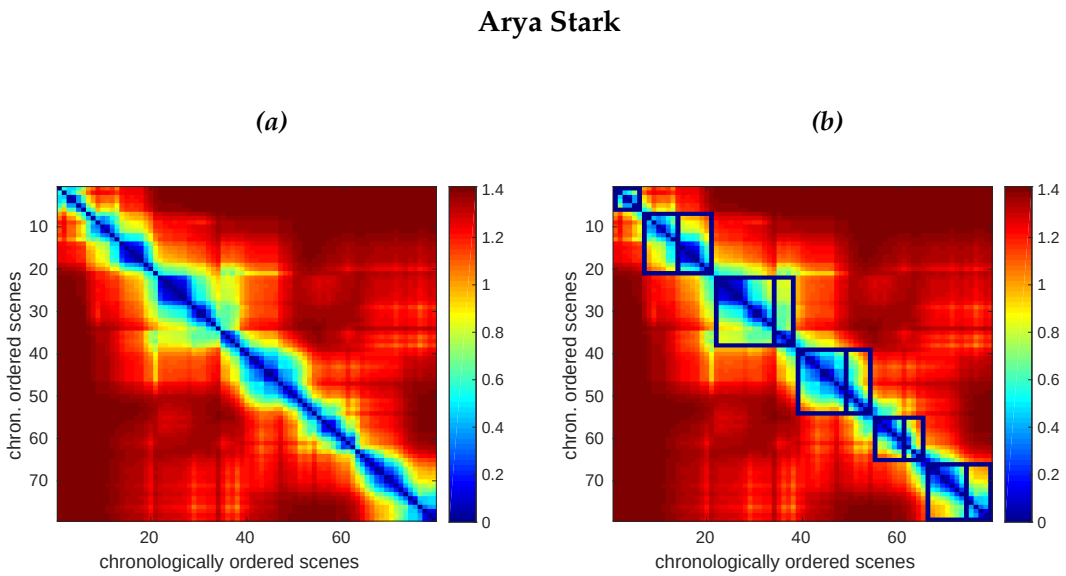


Figure C.1 – *Matrix of distances between Arya Stark's social neighborhoods in any pair of scenes* (t, t') *in the first five seasons of* Game of Thrones*, before (C.1a) and after (C.1b) partitioning with granularity* $\tau := 1.0$

Daenerys Targaryen

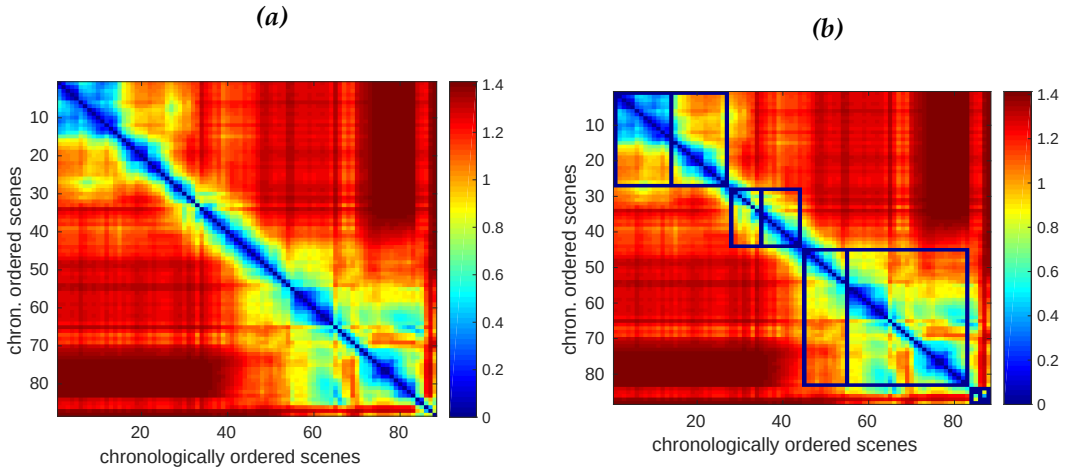


Figure C.2 – *Matrix of distances between Daenerys Targaryen's social neighborhoods in any pair of scenes* (t, t') *in the first five seasons of* Game of Thrones, *before (C.2a) and after (C.2b) partitioning with granularity* $\tau := 1.0$

Jaime Lannister

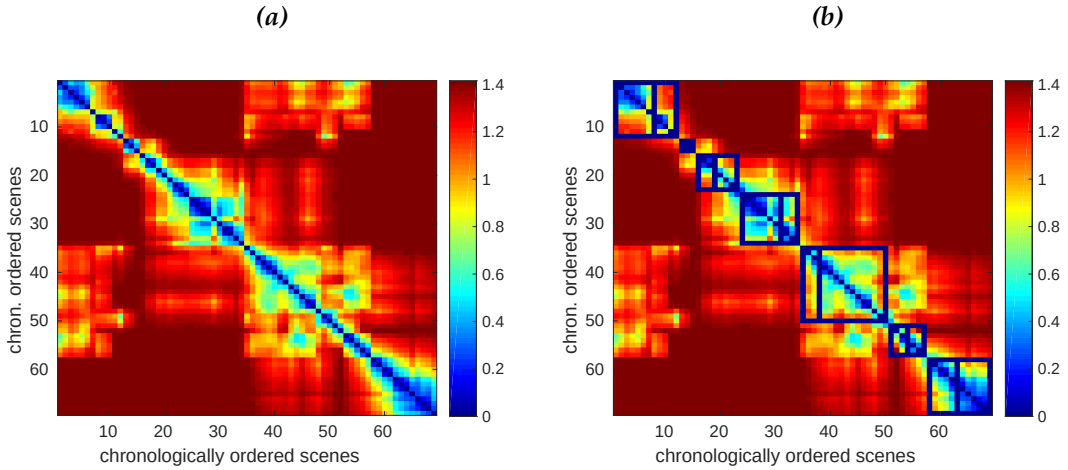


Figure C.3 – *Matrix of distances between Jaime Lannister's social neighborhoods in any pair of scenes* (t, t') *in the first five seasons of* Game of Thrones, *before (C.3a) and after (C.3b) partitioning with granularity* $\tau := 1.0$

Sansa Stark

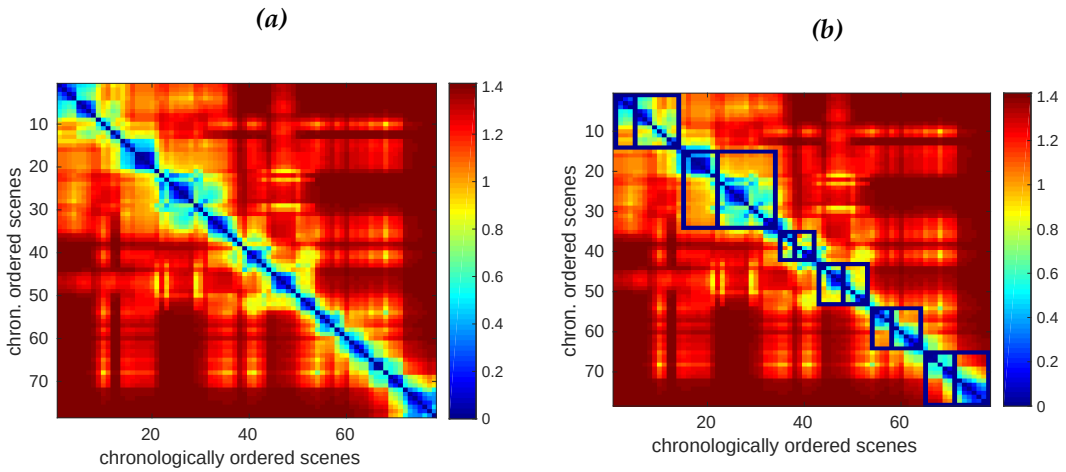


Figure C.4 – *Matrix of distances between Sansa Stark's social neighborhoods in any pair of scenes* (t, t') *in the first five seasons of* Game of Thrones, *before (C.4a) and after (C.4b) partitioning with granularity* $\tau := 1.0$

Theon Greyjoy

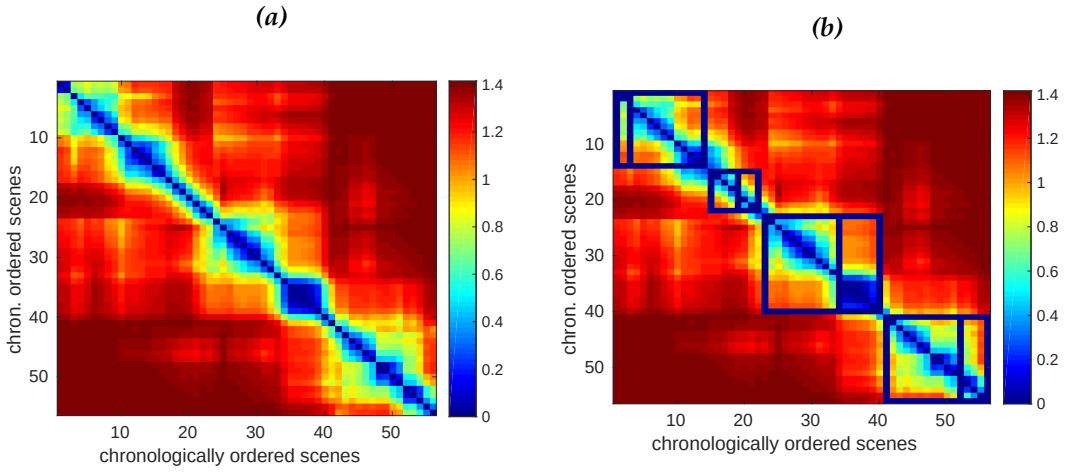


Figure C.5 – *Matrix of distances between Theon Greyjoy's social neighborhoods in any pair of scenes* (t, t') *in the first five seasons of* Game of Thrones, *before (C.5a) and after (C.5b) partitioning with granularity* $\tau := 1.0$

List of Figures

List of Tables

Bibliography

(Agarwal et al., 2012) A. Agarwal, A. Corvalan, J. Jensen, et O. Rambow, 2012. Social network analysis of alice in wonderland. Dans les actes de *Workshop on Computational Linguistics for Literature*, 88–96.

(Agarwal et al., 2013) A. Agarwal, A. Kotalwar, et O. Rambow, 2013. Automatic extraction of social networks from literary text: A case study on alice in wonderland. Dans les actes de *IJCNLP*, 1202–1208.

(Alberich et al., 2002) R. Alberich, J. Miro-Julia, et F. Rosselló, 2002. Marvel universe looks almost like a real social network. *arXiv preprint cond-mat/0202174*.

(Assfalg et al., 2003) J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, et W. Nunziati, 2003. Semantic annotation of soccer videos: automatic highlights identification. *Computer Vision and Image Understanding 92*(2), 285–305.

(Balas et Zemel, 1980) E. Balas et E. Zemel, 1980. An algorithm for large zero-one knapsack problems. *operations Research 28*(5), 1130–1154.

(Bendris et al., 2013) M. Bendris, B. Favre, D. Charlet, G. Damnati, G. Senay, R. Auguste, et J. Martinet, 2013. Unsupervised face identification in tv content using audiovisual sources. Dans les actes de *Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on*, 243–249. IEEE.

(Boreczky et Rowe, 1996) J. S. Boreczky et L. A. Rowe, 1996. Comparison of video shot boundary detection techniques. *Journal of Electronic Imaging 5*(2), 122–128.

(Bost et Linares, 2014) X. Bost et G. Linares, 2014. Constrained speaker diarization of tv series based on visual patterns. Dans les actes de *Spoken Language Technology Workshop (SLT), 2014 IEEE*, 390–395. IEEE.

(Bost et al., 2015) X. Bost, G. Linares, et S. Gueye, 2015. Audiovisual speaker diarization of tv series. Dans les actes de *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 4799–4803. IEEE.

(Bousquet et al., 2011) P.-M. Bousquet, D. Matrouf, et J.-F. Bonastre, 2011. Intersession compensation and scoring methods in the i-vectors space for speaker recognition. Dans les actes de *INTERSPEECH*, 485–488.

(Bozonnet et al., 2010) S. Bozonnet, N. W. Evans, et C. Fredouille, 2010. The lia-eurecom rt'09 speaker diarization system: enhancements in speaker modelling and cluster purification. Dans les actes de *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 4958–4961. IEEE.

(Bredin, 2012) H. Bredin, 2012. Segmentation of tv shows into scenes using speaker diarization and speech recognition. Dans les actes de *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2377–2380. IEEE.

(Bredin et Poignant, 2013) H. Bredin et J. Poignant, 2013. Integer linear programming for speaker diarization and cross-modal identification in tv broadcast. Dans les actes de *the 14rd Annual Conference of the International Speech Communication Association, INTERSPEECH*.

(Bunke et al., 2007) H. Bunke, P. J. Dickinson, M. Kraetzl, et W. D. Wallis, 2007. *A graph-theoretic approach to enterprise network dynamics*, Volume 24. Springer Science & Business Media.

(Carbonell et Goldstein, 1998) J. Carbonell et J. Goldstein, 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. Dans les actes de *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 335–336. ACM.

(Carletta et al., 2005) J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, et al., 2005. *The AMI meeting corpus: A pre-announcement*. Springer.

(Chang et al., 2002) P. Chang, M. Han, et Y. Gong, 2002. Extract highlights from baseball game video with hidden markov models. Dans les actes de *Image Processing. 2002. Proceedings. 2002 International Conference on*, Volume 1, I–609. IEEE.

(Chen et al., 2004) H.-W. Chen, J.-H. Kuo, W.-T. Chu, et J.-L. Wu, 2004. Action movies segmentation and summarization based on tempo analysis. Dans les actes de *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, 251–258. ACM.

(Clauset et Eagle, 2012) A. Clauset et N. Eagle, 2012. Persistence and periodicity in a dynamic proximity network. *arXiv preprint arXiv:1211.7343*.

(Clément et al., 2011) P. Clément, T. Bazillon, et C. Fredouille, 2011. Speaker diarization of heterogeneous web video files: A preliminary study. Dans les actes de *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 4432–4435. IEEE.

(Combes, 2013) C. Combes, 2013. *La pratique des séries télévisées: une sociologie de l'activité spectatorielle*. Thèse de Doctorat, Ecole Nationale Supérieure des Mines de Paris.

(Cour et al., 2008) T. Cour, C. Jordan, E. Miltsakaki, et B. Taskar, 2008. Movie/script: Alignment and parsing of video and text transcription. Dans les actes de *European Conference on Computer Vision*, 158–171. Springer.

(Daskin, 2011) M. S. Daskin, 2011. *Network and discrete location: models, algorithms, and applications*. John Wiley & Sons.

(Davidson et Ravi, 2009) I. Davidson et S. Ravi, 2009. Using instance-level constraints in agglomerative hierarchical clustering: theoretical and empirical results. *Data mining and knowledge discovery 18*(2), 257–282.

(Dehak et al., 2011) N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, et P. Ouellet, 2011. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on 19*(4), 788–798.

(Dupuy et al., 2014) G. Dupuy, S. Meignier, P. Deléglise, et Y. Esteve, 2014. Recent improvements on ilp-based clustering for broadcast news speaker diarization. Dans les actes de *Proc. Odyssey Workshop*.

(Dupuy et al., 2012) G. Dupuy, M. Rouvier, S. Meignier, et Y. Esteve, 2012. I-vectors and ilp clustering adapted to cross-show speaker diarization. Dans les actes de *INTERSPEECH*, 2174–2177.

(Edmundson, 1969) H. P. Edmundson, 1969. New methods in automatic extracting. *Journal of the ACM (JACM) 16*(2), 264–285.

(Elson et al., 2010) D. K. Elson, N. Dames, et K. R. McKeown, 2010. Extracting social networks from literary fiction. Dans les actes de *Proceedings of the 48th annual meeting of the association for computational linguistics*, 138–147. Association for Computational Linguistics.

(Ercolessi, 2013) P. Ercolessi, 2013. *Extraction multimodale de la structure narrative des épisodes de séries télévisées*. Thèse de Doctorat, Université de Toulouse, Université Toulouse III-Paul Sabatier.

(Ercolessi et al., 2011) P. Ercolessi, H. Bredin, C. Sénac, et P. Joly, 2011. Segmenting tv series into scenes using speaker diarization. Dans les actes de *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2011), Delft-Pays bas*, 13–15.

(Ercolessi et al., 2012) P. Ercolessi, C. Sénac, et H. Bredin, 2012. Toward plot deinterlacing in tv series using scenes clustering. Dans les actes de *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, 1–6. IEEE.

(Ethis, 2006) E. Ethis, 2006. *Les spectateurs du temps: pour une sociologie de la réception du cinéma*. Editions L'Harmattan.

(Ethis et Fabiani, 2004) E. Ethis et J.-L. Fabiani, 2004. Pour une po (i) étique du questionnaire en sociologie de la culture : le spectateur imaginé. *Logiques sociales*.

(Evangelopoulos et al., 2013) G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, et Y. Avrithis, 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia 15*(7), 1553–1568.

(Evans et al., 2012) N. Evans, S. Bozonnet, D. Wang, C. Fredouille, et R. Troncy, 2012. A comparative study of bottom-up and top-down approaches to speaker diarization. *Audio, Speech, and Language Processing, IEEE Transactions on 20*(2), 382–392.

(Fayard et Plateau, 1982) D. Fayard et G. Plateau, 1982. An algorithm for the solution of the 0–1 knapsack problem. *Computing 28*(3), 269–287.

(Felzenszwalb et al., 2008) P. Felzenszwalb, D. McAllester, et D. Ramanan, 2008. A discriminatively trained, multiscale, deformable part model. Dans les actes de *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

(Ferrari et al., 2009) V. Ferrari, M. Marin-Jimenez, et A. Zisserman, 2009. Pose search: retrieving people using their pose. Dans les actes de *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1–8. IEEE.

(Friedland et al., 2009a) G. Friedland, L. Gottlieb, et A. Janin, 2009a. Using artistic markers and speaker identification for narrative-theme navigation of seinfeld episodes. Dans les actes de *Multimedia, 2009. ISM'09. 11th IEEE International Symposium on*, 511–516. IEEE.

(Friedland et al., 2009b) G. Friedland, H. Hung, et C. Yeo, 2009b. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. Dans les actes de *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 4069–4072. IEEE.

(Giannakopoulos et al., 2008) T. Giannakopoulos, A. Pikrakis, et S. Theodoridis, 2008. Music tracking in audio streams from movies. Dans les actes de *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, 950–955. IEEE.

(Gillick et Favre, 2009) D. Gillick et B. Favre, 2009. A scalable global model for summarization. Dans les actes de *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, 10–18. Association for Computational Linguistics.

(Girshick et al., 2012) R. B. Girshick, P. F. Felzenszwalb, et D. McAllester, 2012. Discriminatively trained deformable part models, release 5.

(Gleiser, 2007) P. M. Gleiser, 2007. How to become a superhero. *Journal of Statistical Mechanics: Theory and Experiment 2007*(09), P09020.

(Gudivada et Raghavan, 1995) V. N. Gudivada et V. V. Raghavan, 1995. Content based image retrieval systems. *Computer 28*(9), 18–22.

(Guha et al., 2015) T. Guha, N. Kumar, S. S. Narayanan, et S. L. Smith, 2015. Computationally deconstructing movie narratives: an informatics approach. Dans les actes de *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2264–2268. IEEE.

(Hakimi, 1964) S. L. Hakimi, 1964. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations research 12*(3), 450–459.

(Hakimi, 1965) S. L. Hakimi, 1965. Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research 13*(3), 462–475.

(Hanjalic et al., 1999) A. Hanjalic, R. L. Lagendijk, et J. Biemond, 1999. Automated high-level movie segmentation for advanced video-retrieval systems. *Circuits and Systems for Video Technology, IEEE Transactions on 9*(4), 580–588.

(Hanjalic et Xu, 2005) A. Hanjalic et L.-Q. Xu, 2005. Affective video content representation and modeling. *Multimedia, IEEE Transactions on 7*(1), 143–154.

(Holme et Saramäki, 2012) P. Holme et J. Saramäki, 2012. Temporal networks. *Physics reports 519*(3), 97–125.

(Irie et al., 2010) G. Irie, T. Satou, A. Kojima, T. Yamasaki, et K. Aizawa, 2010. Automatic trailer generation. Dans les actes de *Proceedings of the 18th ACM international conference on Multimedia*, 839–842. ACM.

(Kaminski et al., 2012) J. Kaminski, M. Schober, R. Albaladejo, O. Zastupailo, et C. Hidalgo, 2012. Moviegalaxies - social networks in movies.

(Klastorin, 1985) T. D. Klastorin, 1985. The p-median problem for cluster analysis: a comparative test using the mixture model approach. *Management Science 31*(1), 84–95.

(Koprinska et Carrato, 2001) I. Koprinska et S. Carrato, 2001. Temporal video segmentation: A survey. *Signal processing: Image communication 16*(5), 477–500.

(Larcher et al., 2013) A. Larcher, J.-F. Bonastre, B. G. Fauve, K.-A. Lee, C. Lévy, H. Li, J. S. Mason, et J.-Y. Parfait, 2013. Alize 3.0-open source toolkit for state-of-the-art speaker recognition. Dans les actes de *Interspeech*, 2768–2772.

(Lartillot et al., 2008) O. Lartillot, P. Toiviainen, et T. Eerola, 2008. A matlab toolbox for music information retrieval. Dans les actes de *Data analysis, machine learning and applications*, 261–268. Springer.

(Li et Sezan, 2001) B. Li et M. I. Sezan, 2001. Event detection and summarization in sports video. Dans les actes de *Content-Based Access of Image and Video Libraries, 2001.(CBAIVL 2001). IEEE Workshop on*, 132–138. IEEE.

(Liu et al., 2013) S. Liu, Y. Wu, E. Wei, M. Liu, et Y. Liu, 2013. Storyflow: Tracking the evolution of stories. *IEEE Transactions on Visualization and Computer Graphics 19*(12), 2436–2445.

(Ma et al., 2002) Y.-F. Ma, L. Lu, H.-J. Zhang, et M. Li, 2002. A user attention model for video summarization. Dans les actes de *Proceedings of the tenth ACM international conference on Multimedia*, 533–542. ACM.

(Mac Carron et Kenna, 2012) P. Mac Carron et R. Kenna, 2012. Universal properties of mythological networks. *EPL (Europhysics Letters) 99*(2), 28002.

(Malinas, 2015) D. Malinas, 2015. Démocratisation culturelle et numérique. *Culture et Musées* (24).

(McDonald, 2007) R. McDonald, 2007. *A study of global inference algorithms in multi-document summarization*. Springer.

(Meignier et Merlin, 2010) S. Meignier et T. Merlin, 2010. Lium spkdiarization: an open source toolkit for diarization. Dans les actes de *CMU SPUD Workshop*, Volume 2010.

(Min et Park, 2016) S. Min et J. Park, 2016. Mapping out narrative structures and dynamics using networks and textual information. *arXiv cs.CL*, 1604.03029.

(Monaco, 2000) J. Monaco, 2000. *How to read a film: the world of movies, media, and multimedia: language, history, theory*. Oxford University Press, USA.

(Money et Agius, 2008) A. G. Money et H. Agius, 2008. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation 19*(2), 121–143.

(Moretti, 2011) F. Moretti, 2011. Network theory, plot analysis. *New Left Review*.

(Mutton, 2004) P. Mutton, 2004. Inferring and visualizing social networks on internet relay chat. Dans les actes de *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*, 35–43. IEEE.

(Nalisnick et Baird, 2013) E. T. Nalisnick et H. S. Baird, 2013. Character-to-character sentiment analysis in Shakespeare's plays. Dans les actes de *51st Annual Meeting of the Association for Computational Linguistics*.

(Nemhauser et Wolsey, 1988) G. L. Nemhauser et L. A. Wolsey, 1988. Integer and combinatorial optimization. interscience series in discrete mathematics and optimization. *ed: John Wiley & Sons*.

(Ogawa et Ma, 2010) M. Ogawa et K.-L. Ma, 2010. Software evolution storylines. Dans les actes de *Proceedings of the 5th international symposium on Software visualization*, 35–42. ACM.

(Rochat et Kaplan, 2014) Y. Rochat et F. Kaplan, 2014. Analyse de réseaux sur les personnages des confessions de Jean-Jacques Rousseau. *Cahiers du numérique 10*(3), 109–133.

(Roth, 2013) R. Roth, 2013. *Bande originale de film, bande originale de vie: pour une sémiologie tripartite de l'emblème musical: le cas de l'univers Disney*. Thèse de Doctorat, Université d'Avignon.

(Rousseeuw, 1987) P. J. Rousseeuw, 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20*, 53–65.

(Rouvier et al., 2013) M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, et S. Meignier, 2013. An open-source state-of-the-art toolbox for broadcast news diarization. Rapport technique, Idiap.

(Roy et al., 2014) A. Roy, C. Guinaudeau, H. Bredin, et C. Barras, 2014. Tvd: A reproducible and multiply aligned tv series dataset. Dans les actes de *LREC*, 418–425.

(Sang et Xu, 2010) J. Sang et C. Xu, 2010. Character-based movie summarization. Dans les actes de *Proceedings of the international conference on Multimedia*, 855–858. ACM.

(Sivic et al., 2009) J. Sivic, M. Everingham, et A. Zisserman, 2009. Who are you? learning person specific classifiers from video. Dans les actes de *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1145–1152. IEEE.

(Smeaton et al., 2006) A. F. Smeaton, B. Lehane, N. E. O'Connor, C. Brady, et G. Craig, 2006. Automatically selecting shots for action movie trailers. Dans les actes de *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, 231–238. ACM.

(Smith et Kanade, 1997) M. A. Smith et T. Kanade, 1997. Video skimming and characterization through the combination of image and language understanding techniques. Dans les actes de *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 775–781. IEEE.

(Sparavigna, 2013) A. C. Sparavigna, 2013. On social networks in plays and novels. *International Journal of Sciences 2*(10), 20–25.

(Tapaswi et al., 2012) M. Tapaswi, M. Bäuml, et R. Stiefelhagen, 2012. Knock! knock! who is it? probabilistic person identification in tv-series. Dans les actes de *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2658–2665. IEEE.

(Tapaswi et al., 2014) M. Tapaswi, M. Bauml, et R. Stiefelhagen, 2014. Storygraphs: visualizing character interactions as a timeline. Dans les actes de *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 827–834.

(Tran et al., 2011) V.-A. Tran, V. B. Le, C. Barras, et L. Lamel, 2011. Comparing multistage approaches for cross-show speaker diarization. Dans les actes de *INTERSPEECH*, Volume 201, 1053–1056.

(Truong et Venkatesh, 2007) B. T. Truong et S. Venkatesh, 2007. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 3*(1), 3.

(Tsai et al., 2013) C.-M. Tsai, L.-W. Kang, C.-W. Lin, et W. Lin, 2013. Scene-based movie summarization via role-community networks. *IEEE Transactions on Circuits and Systems for Video Technology 23*(11), 1927–1940.

(Tsoneva et al., 2007) T. Tsoneva, M. Barbieri, et H. Weda, 2007. Automated summarization of narrative video on a semantic level. Dans les actes de *Semantic Computing, 2007. ICSC 2007. International Conference on*, 169–176. IEEE.

(Tsoumakas et Katakis, 2006) G. Tsoumakas et I. Katakis, 2006. Multi-label classification: An overview. *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*.

(Vendrig et Worring, 2002) J. Vendrig et M. Worring, 2002. Systematic evaluation of logical story unit segmentation. *Multimedia, IEEE Transactions on 4*(4), 492–499.

(Weng et al., 2007) C.-Y. Weng, W.-T. Chu, et J.-L. Wu, 2007. Movie analysis based on roles' social network. Dans les actes de *Multimedia and Expo, 2007 IEEE International Conference on*, 1403–1406. IEEE.

(Weng et al., 2009) C.-Y. Weng, W.-T. Chu, et J.-L. Wu, 2009. Rolenet: Movie analysis from the perspective of social networks. *Multimedia, IEEE Transactions on 11*(2), 256–271.

(Yeung et al., 1998) M. Yeung, B.-L. Yeo, et B. Liu, 1998. Segmentation of video by clustering and graph analysis. *Computer vision and image understanding 71*(1), 94–109.