Users Are Known by the Company They Keep: Topic Models for Viewpoint Discovery in Social Networks

Thibaut THONET Guillaume CABANAC Mohand BOUGHANEM Karen PINEL-SAUVAGNAT

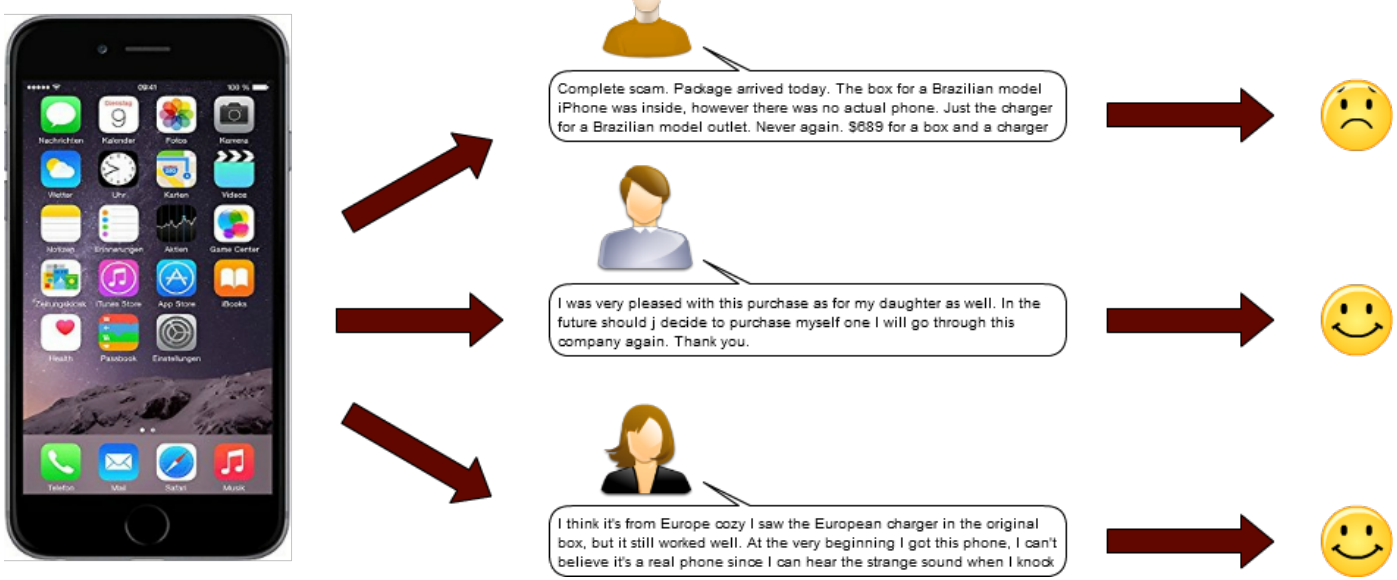
IRIT, Université de Toulouse, CNRS

7 November 2017

Motivation

- Massive amount of **opinions** on the Web
  $\Longrightarrow$ Need for **automated** methods to identify, classify and summarize opinions

- Traditional **opinion mining** research mainly focused on product/service review analysis
  $\Longrightarrow$ Identification of a review's **polarity** w.r.t. a target: **positive**/**negative**



Images and reviews taken from Wikipedia and Amazon.com, February 2016.

Motivation

. . . But **need** to go beyond plain positive/negative opinions $\Longrightarrow$ **viewpoint-based opinions**

E.g., to deal with **filter bubbles** [Pariser, 2011] & **echo chambers** [Sunstein, 2009]

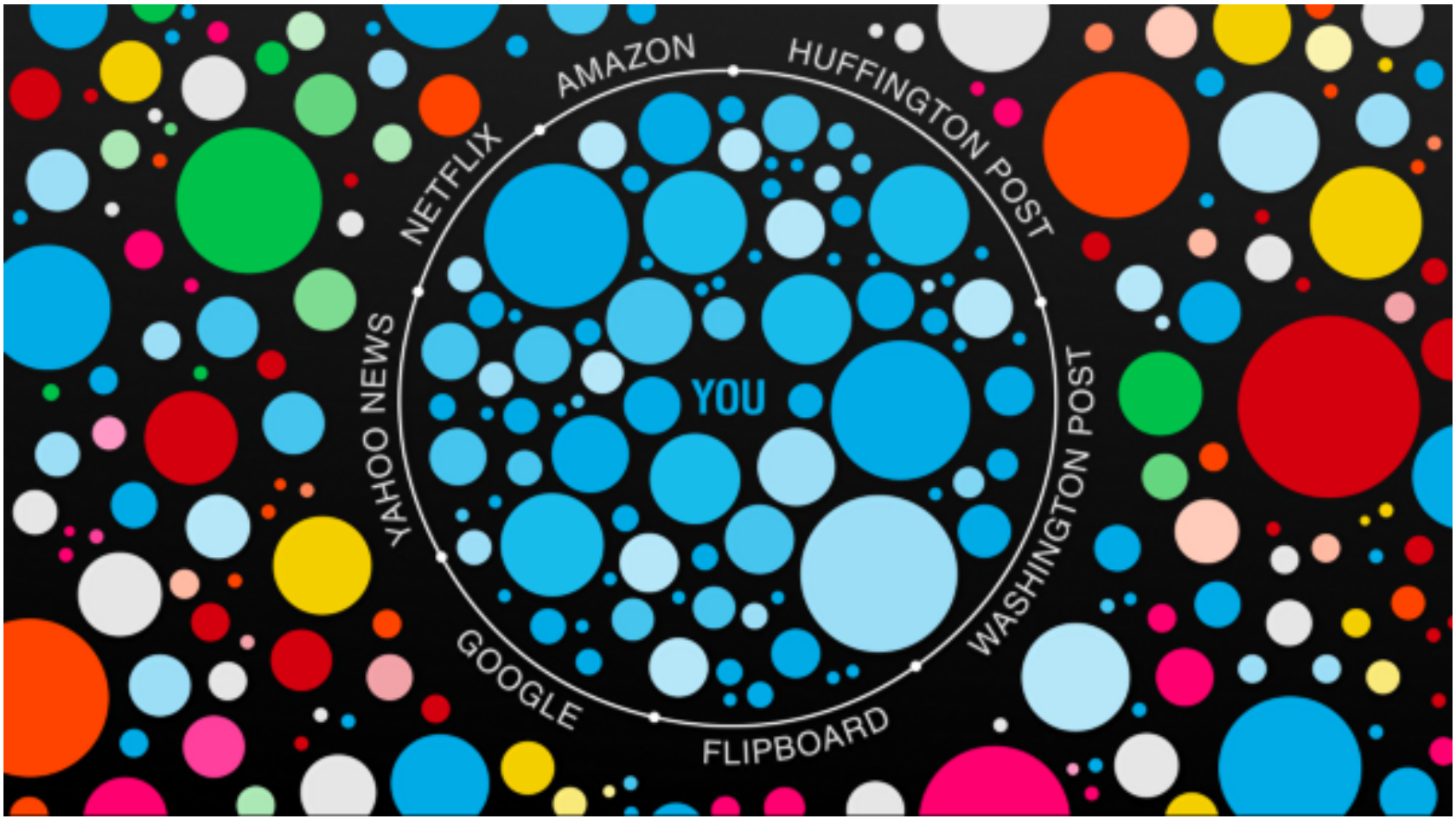
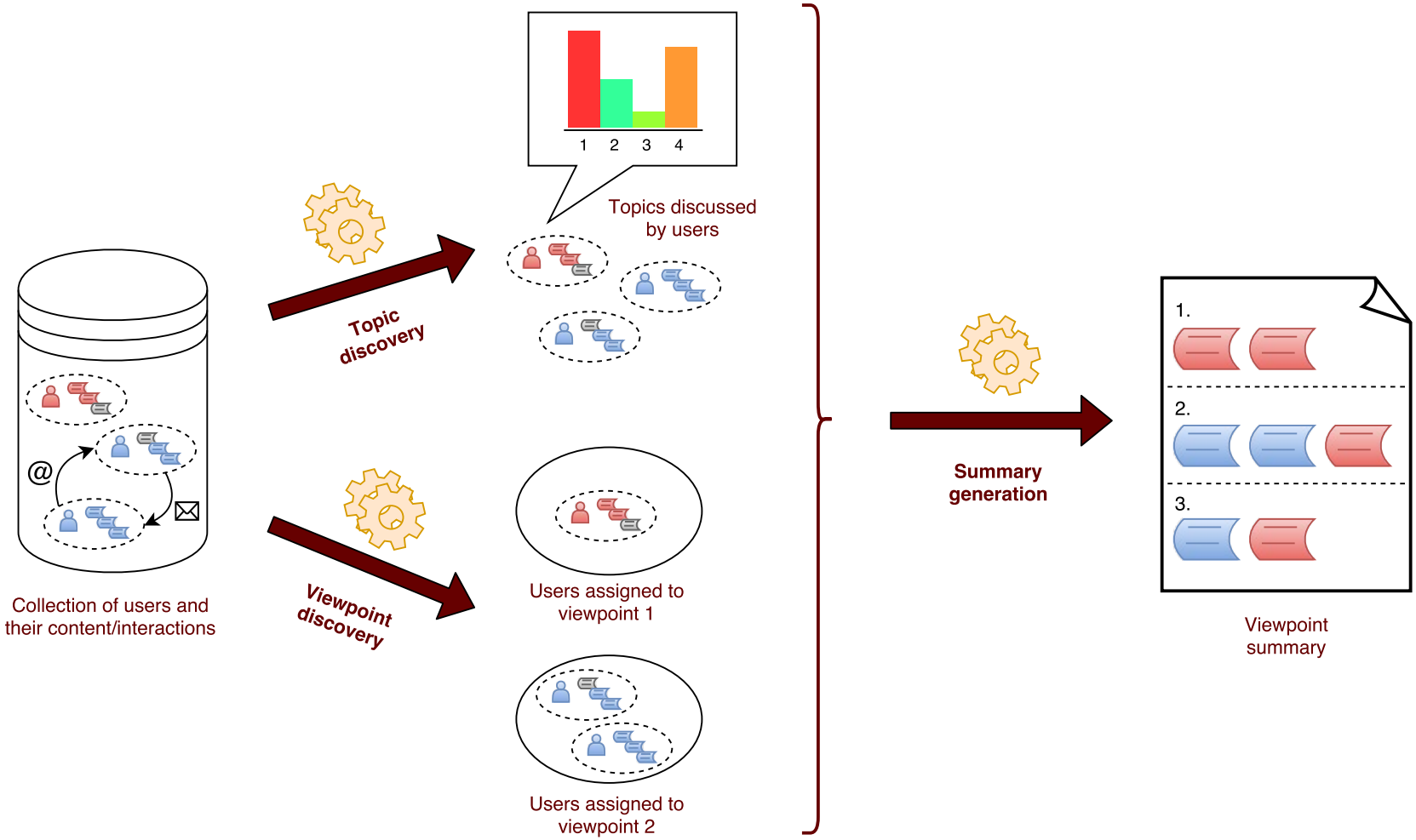


Image taken from wired.com, October 2017.

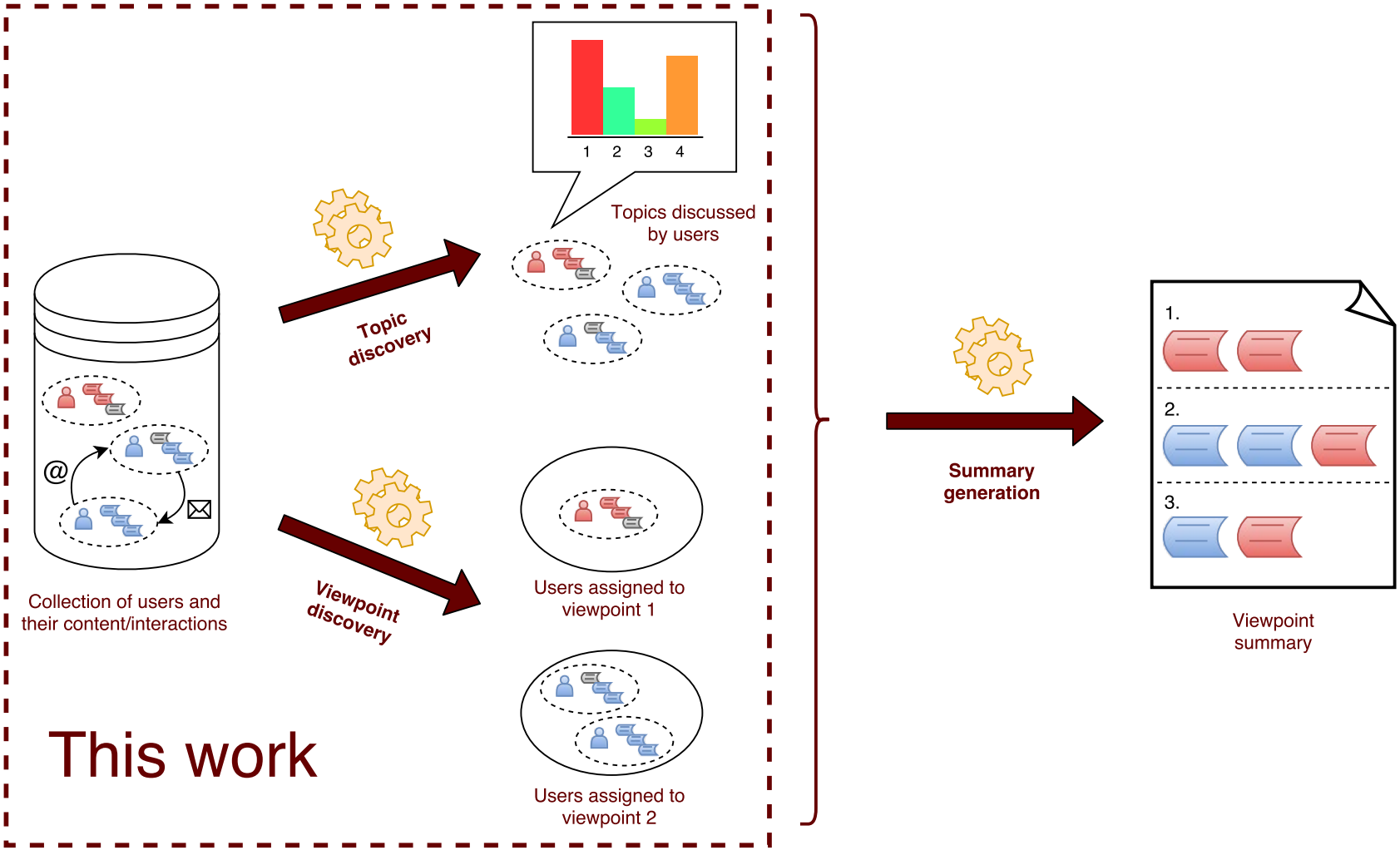
Task

How to mitigate filter bubbles & echo chambers? $\Longrightarrow$ Build **unbiased viewpoint summaries**



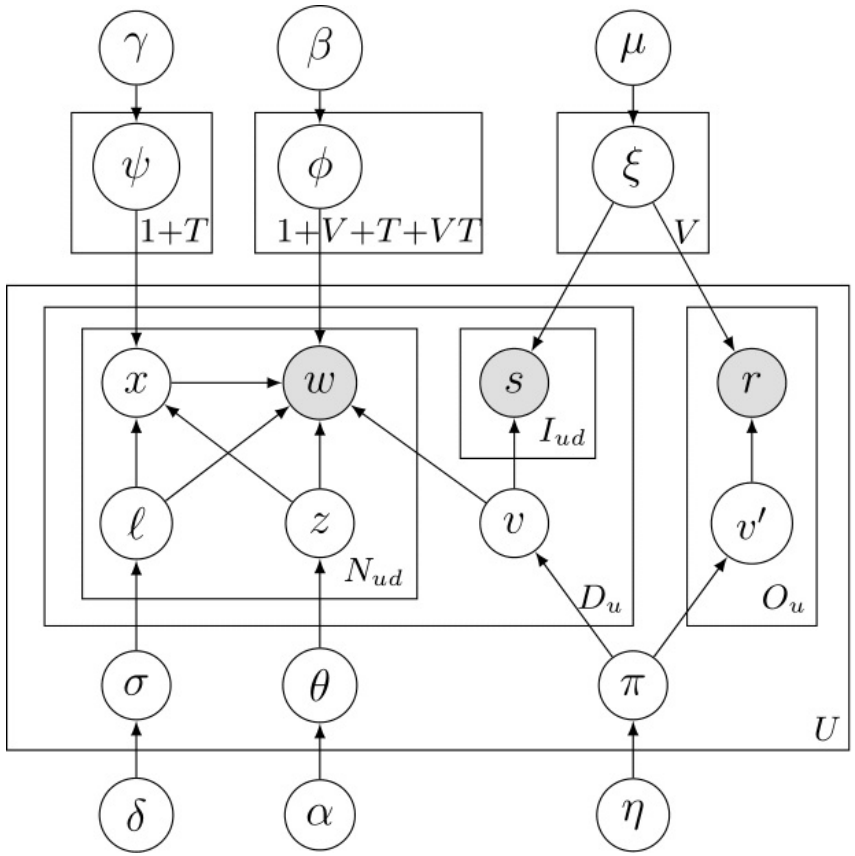
Task

This work is the **first step**: discover **viewpoints** and **topics** from social networking data



SNVDM: The Social Network Viewpoint Discovery Model

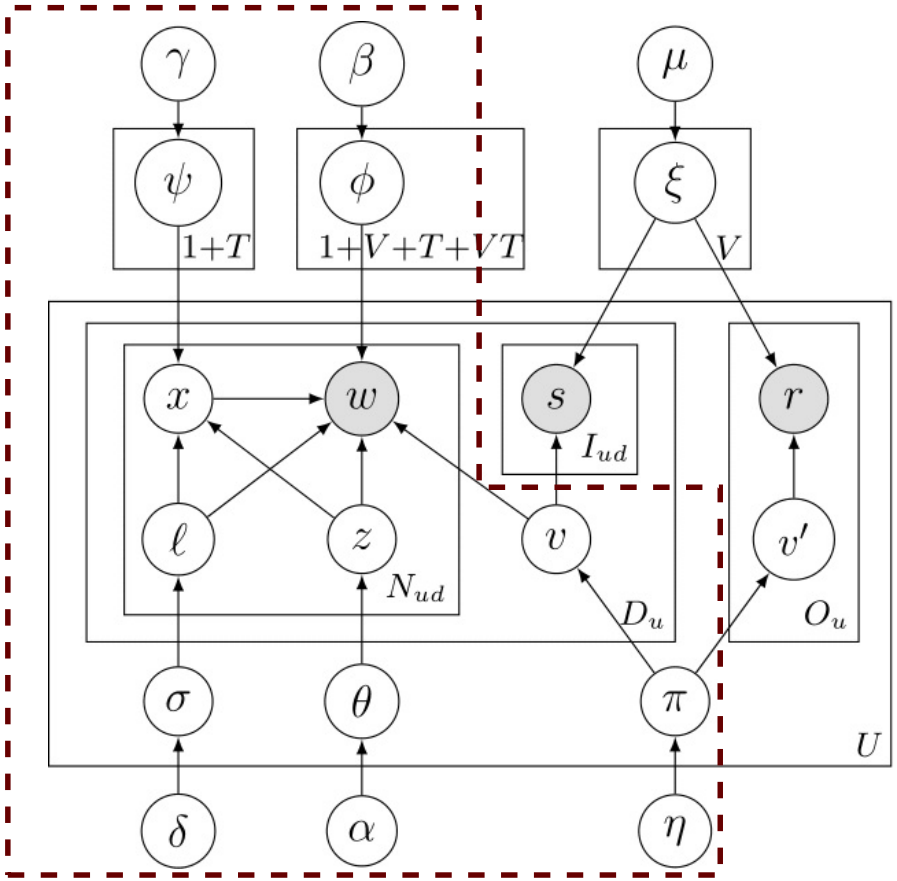
We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**



SNVDM: The Social Network Viewpoint Discovery Model

We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**

Text content component

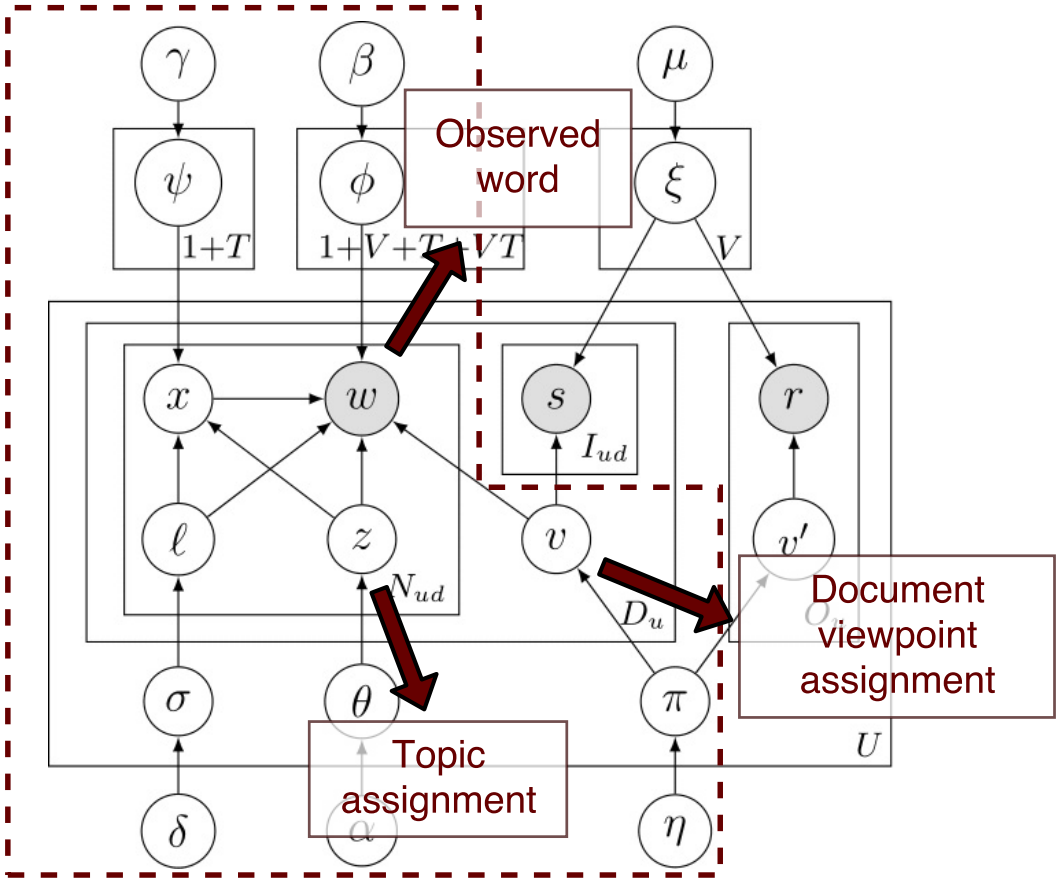


SNVDM: The Social Network Viewpoint Discovery Model

We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**

Text content component

- Observed data: **tokens** occurring in **documents** posted by **users** $\Longrightarrow$ 3 nested plates
- Latent **topics** assigned to each token
- Latent **viewpoints** assigned at **document-level**



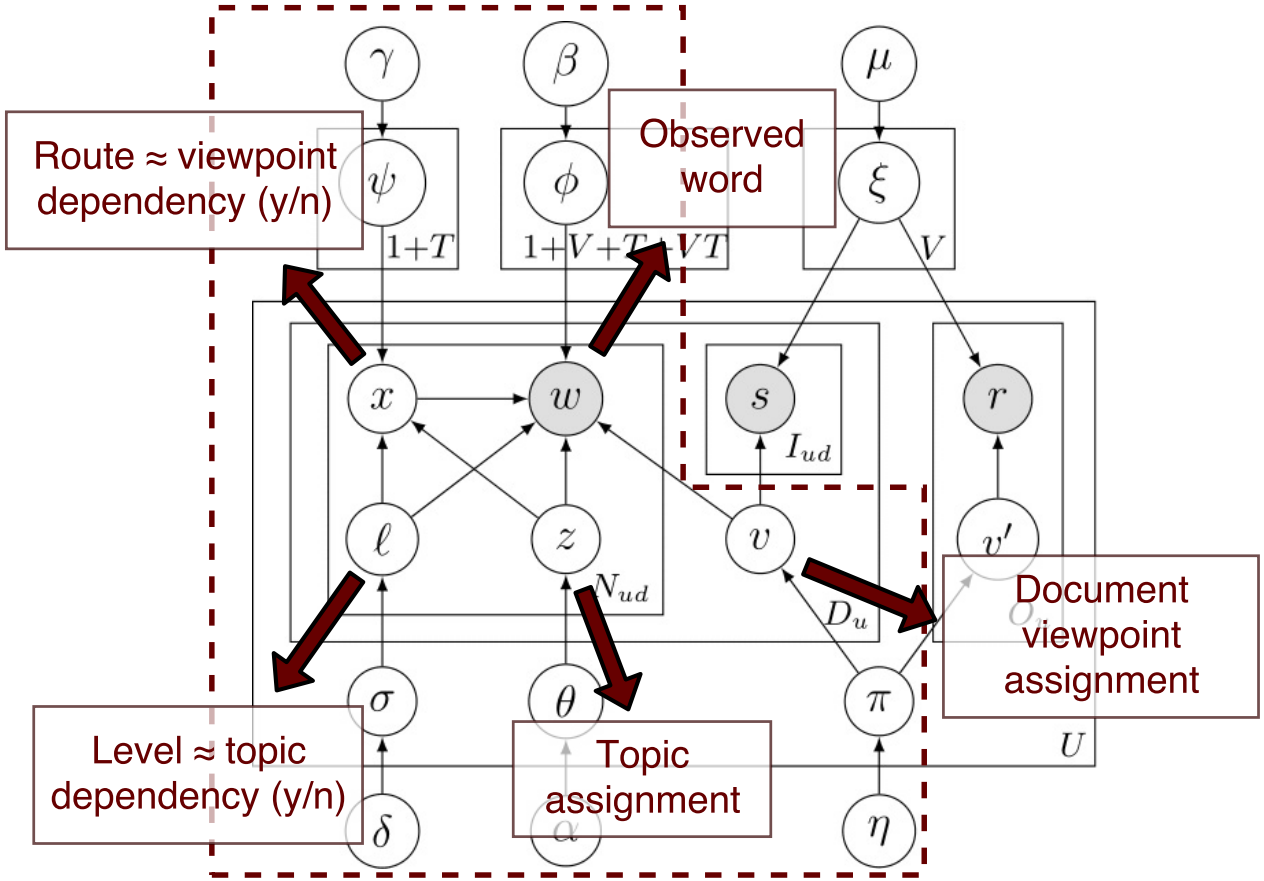
SNVDM: The Social Network Viewpoint Discovery Model

We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**

Text content component

Following the **Topic-Aspect Model** from [Paul+, AAAI '10], definition of **four word types** specified by switch variables ℓ (level) and x (route):

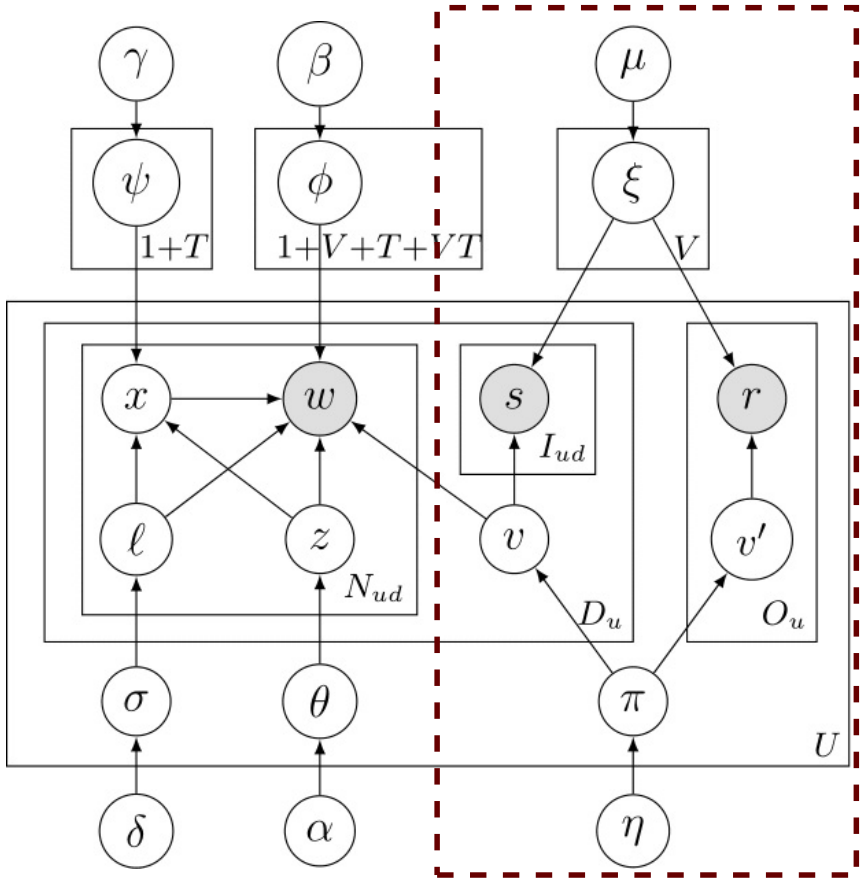
- **Background** words
  $\Longrightarrow \ell = 0, x = 0$
- **Viewpoint** words
  $\Longrightarrow \ell = 0, x = 1$
- **Topic** words
  $\Longrightarrow \ell = 1, x = 0$
- **Viewpoint-topic** words
  $\Longrightarrow \ell = 1, x = 1$



SNVDM: The Social Network Viewpoint Discovery Model

We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**

Social interaction component

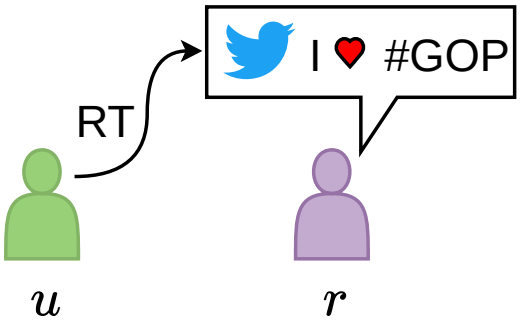


SNVDM: The Social Network Viewpoint Discovery Model

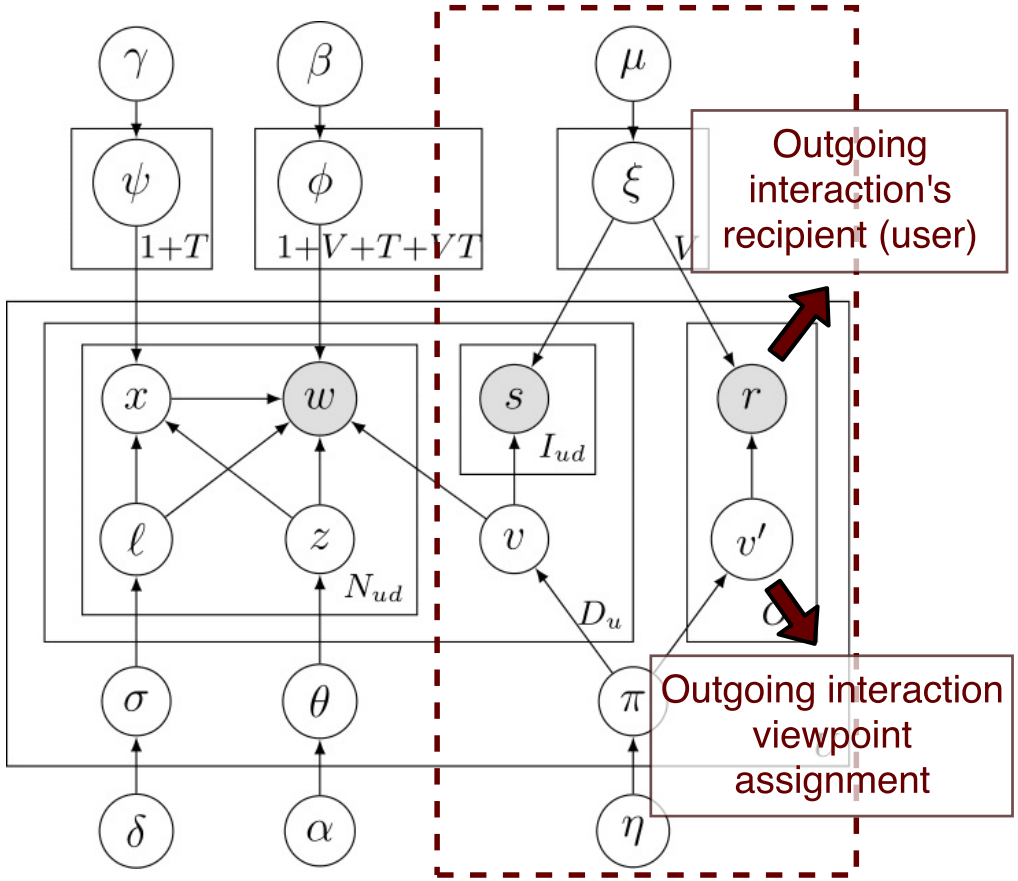
We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**

Social interaction component

Outgoing interactions for user u = interactions initiated by u on another user (**recipient** r)

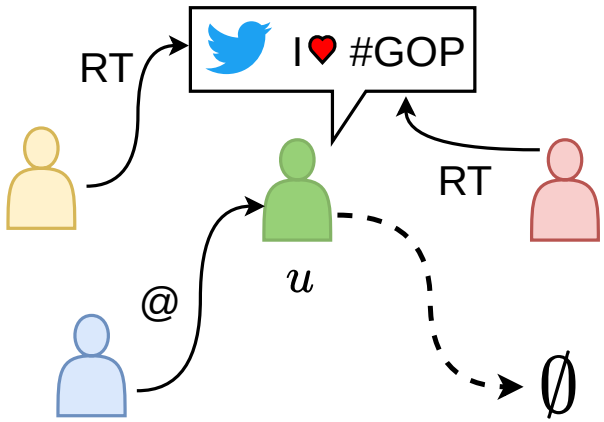


Following **SN-LDA** from [Sachan+, WSDM '14], viewpoints assigned to outgoing interactions (**homophily**)

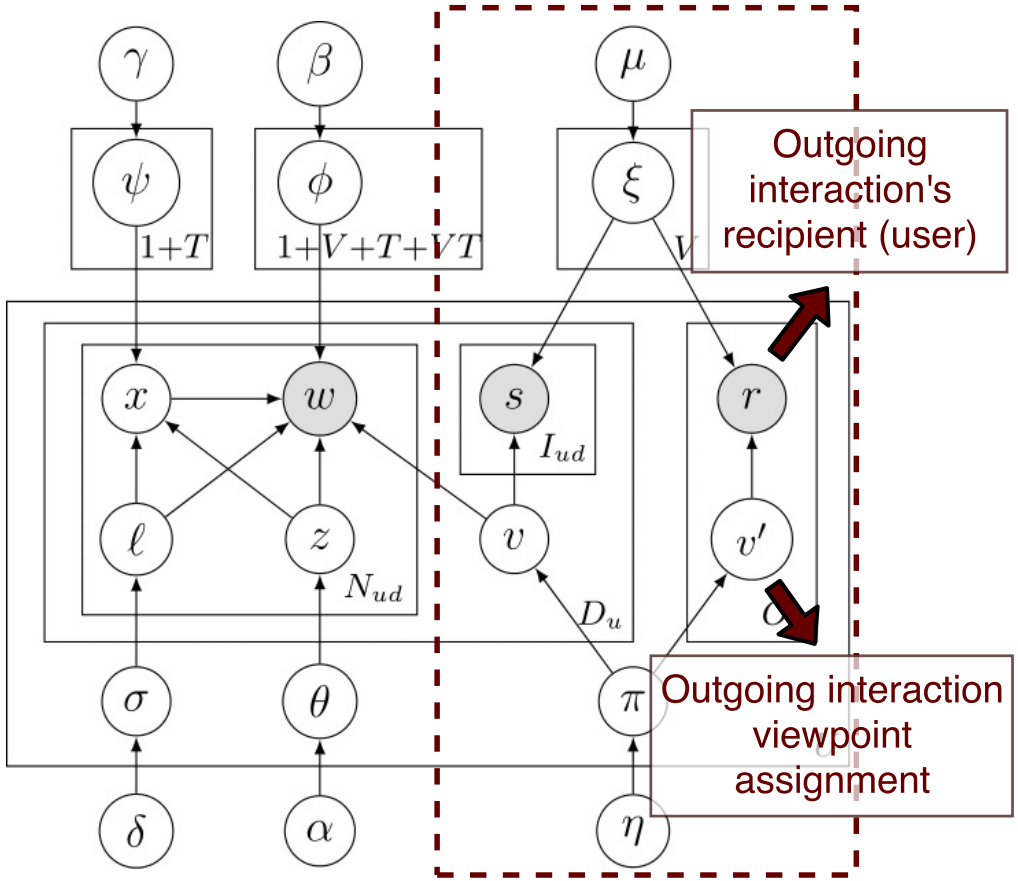


SNVDM: The Social Network Viewpoint Discovery Model

We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**



$\Longrightarrow$ We propose to also exploit **incoming interactions**

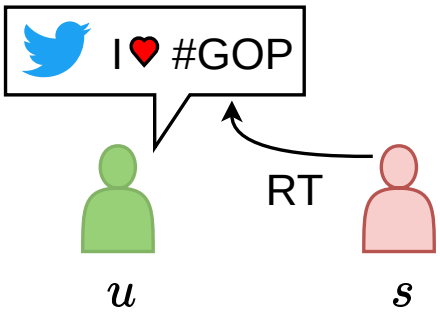


SNVDM: The Social Network Viewpoint Discovery Model

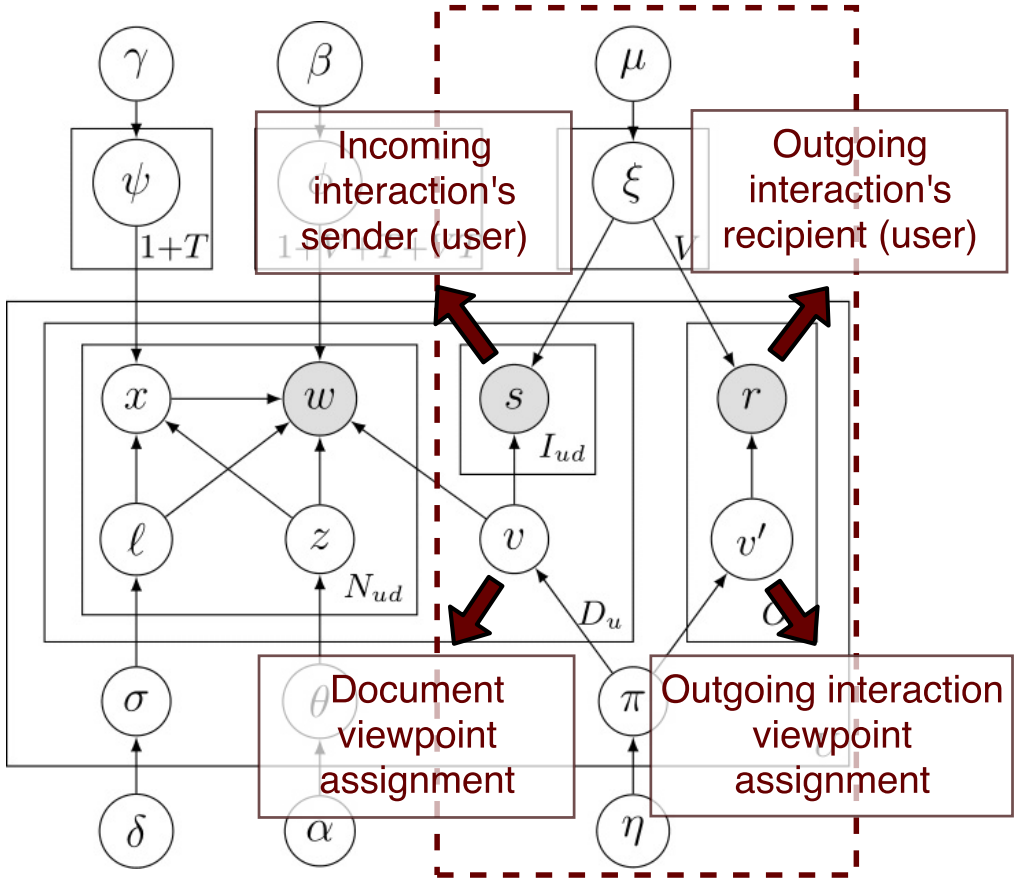
We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**

Social interaction component

Incoming interactions for user u = interactions initiated by another user (**sender** s) on u



Viewpoint assigned to the **document** being interacted upon



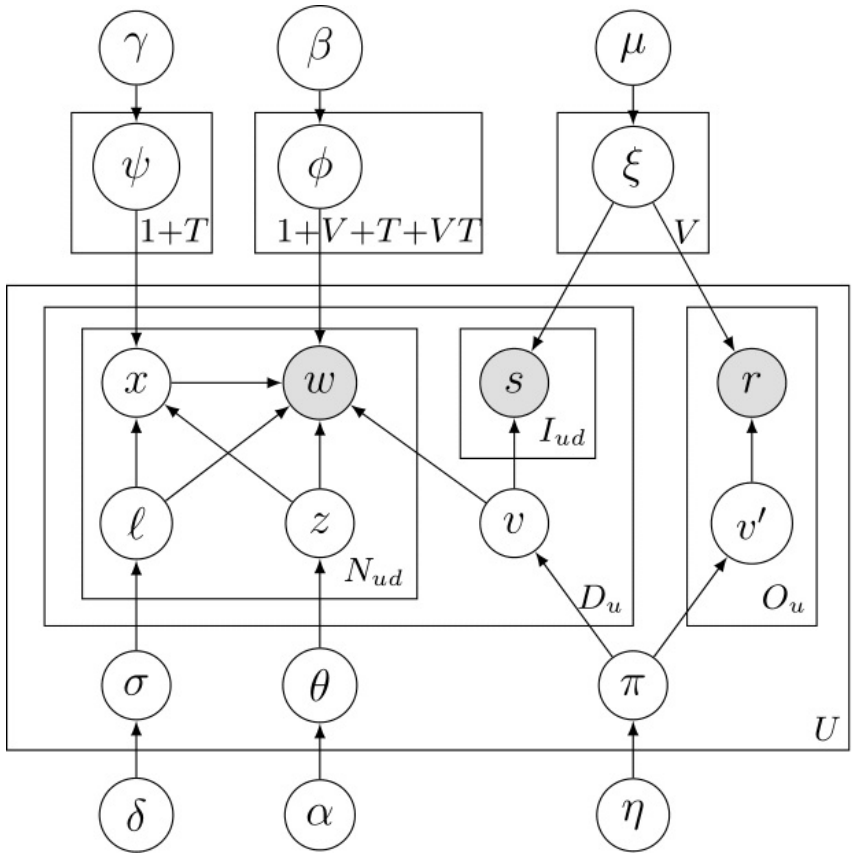
SNVDM: The Social Network Viewpoint Discovery Model

We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**

Posterior inference

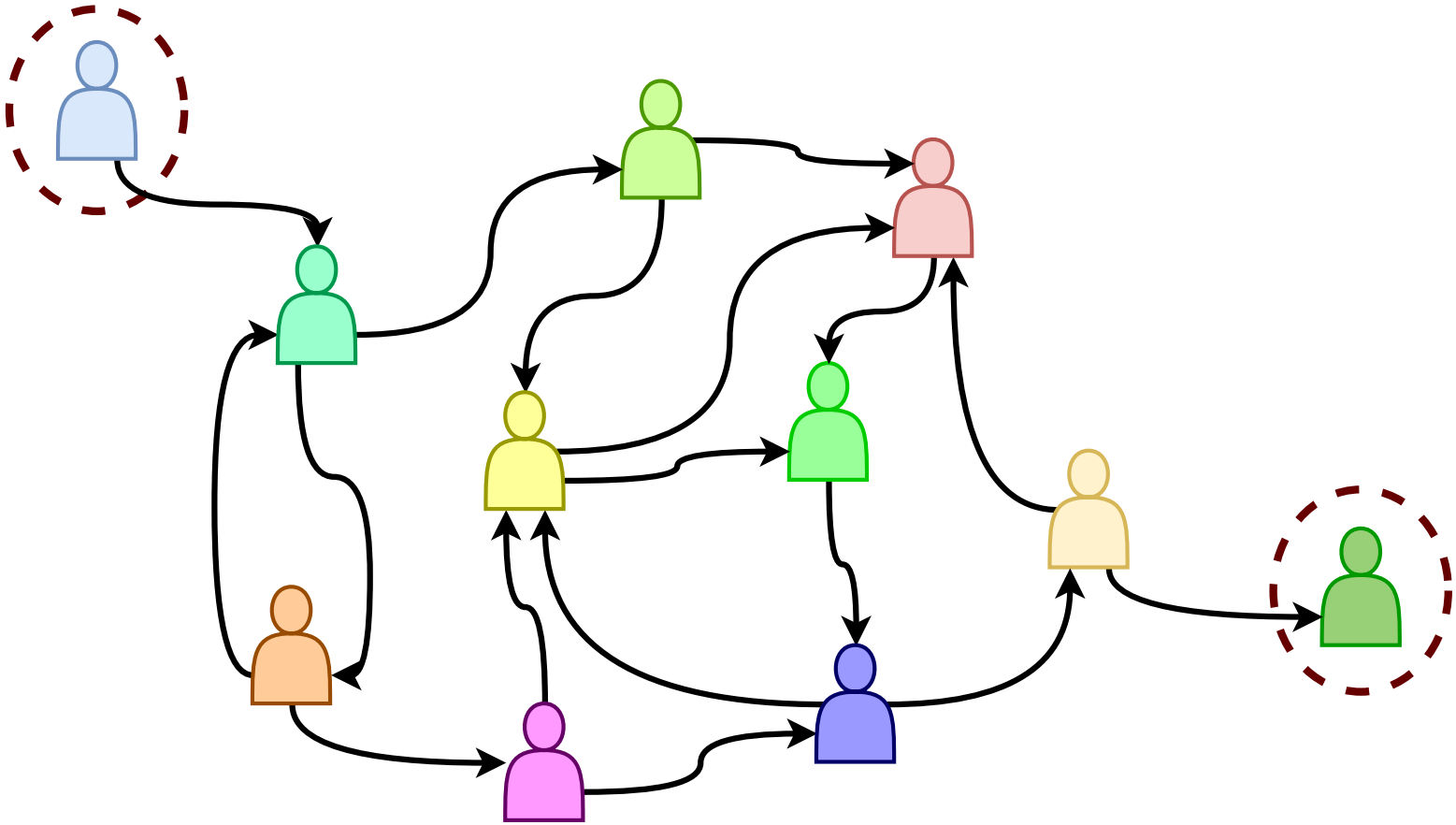
Approximate inference based on **Collapsed Gibbs Sampling**

- Dirichlet/Bernoulli distributions σ, ψ, θ, π, ϕ, ξ **integrated out**
- Successively **sample** discrete latent variables ℓ, x, z, v, v' from their posterior distributions (i.e., given observations w, s, r)



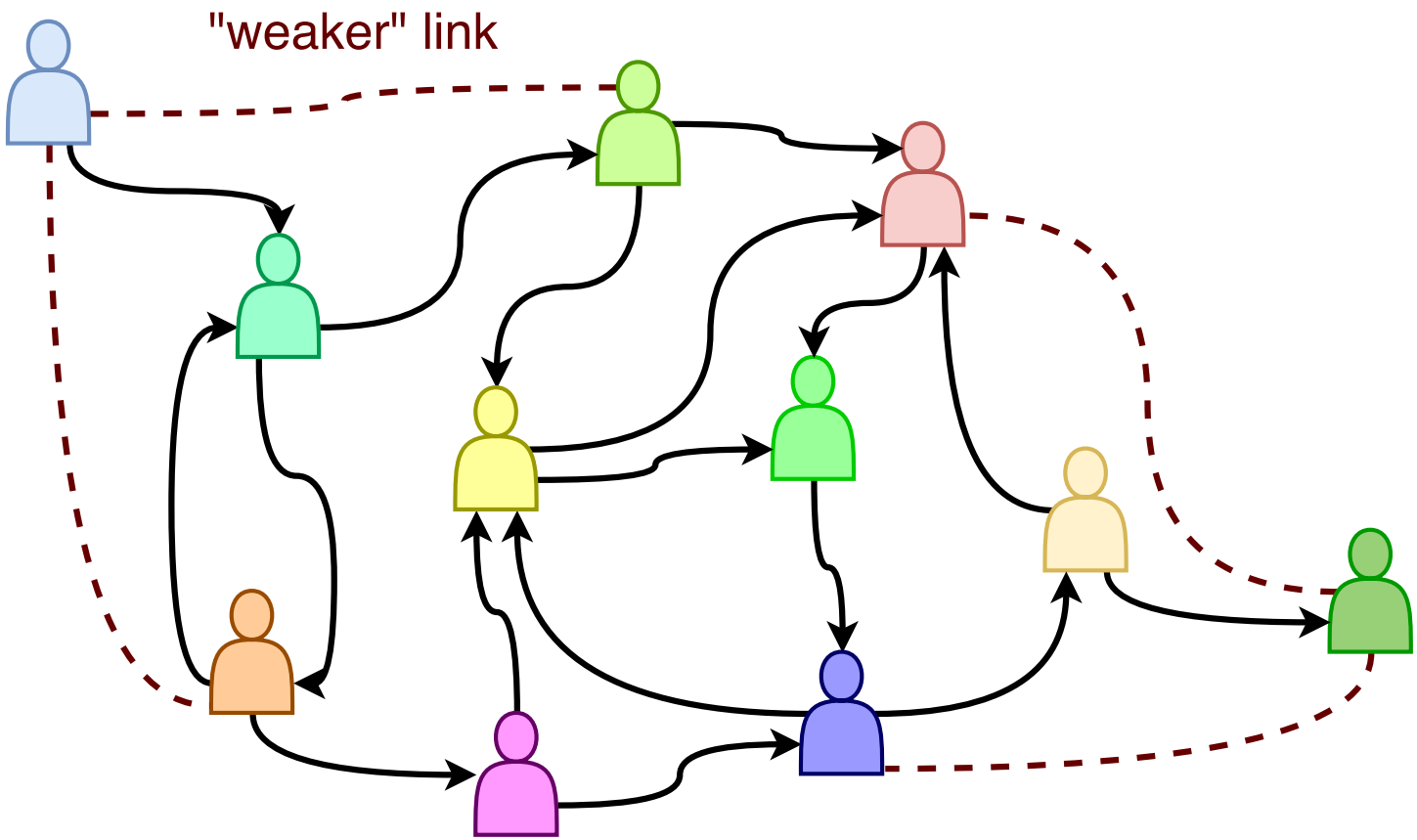
Limits of SNVDM's Social Interaction Component

Some users have **very few social interactions**
$\Longrightarrow$ Difficult to identify their viewpoints based on scarce **direct** interactions



Limits of SNVDM's Social Interaction Component

We propose to extend SNVDM to leverage "aquaintances of acquaintances" ($\approx$ friends of friends)
How? $\Longrightarrow$ **Generalized Pólya Urn** scheme



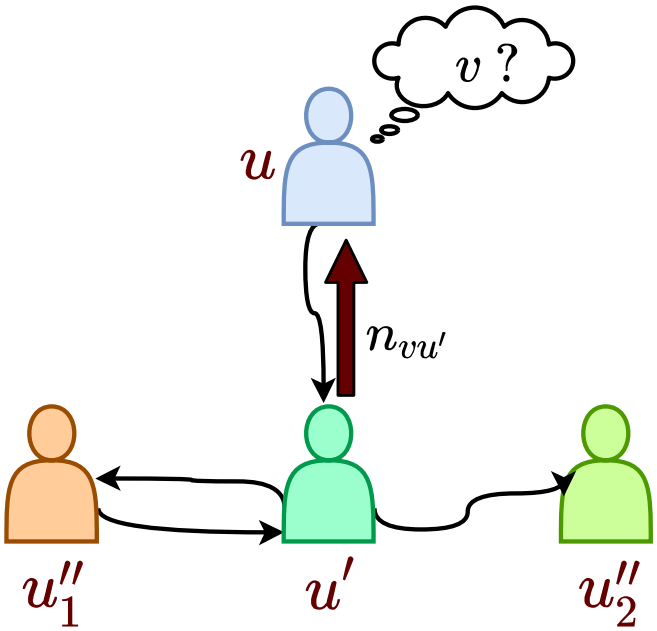
SNVDM-GPU: Extension of SNVDM based on Generalized Pólya Urn

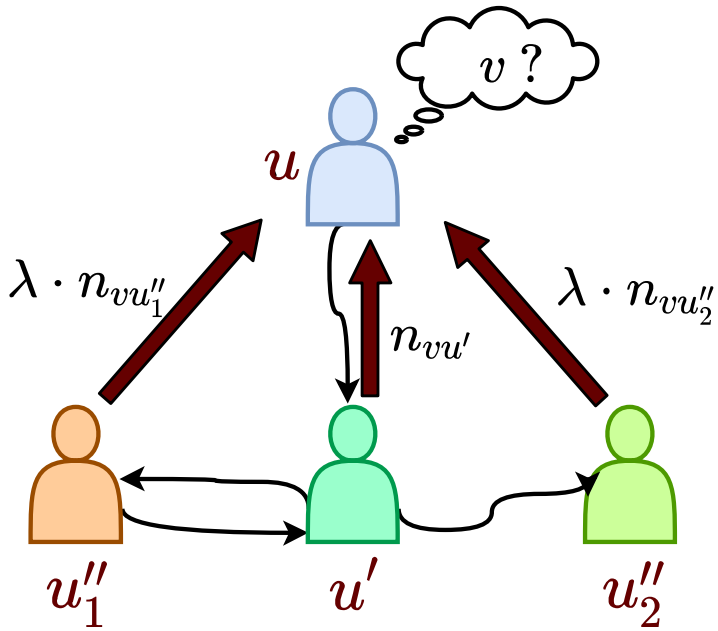
Using **Generalized Pólya Urn** in SNVDM requires minor changes in collapsed Gibbs sampling

E.g., for outgoing interaction o from user u on user u':

$$p(v'_{uo} = v | r_{uo} = u', \text{rest}) \propto \frac{n_{uv} + \eta \frac{1}{V}}{n_{u\cdot} + \eta} \cdot \frac{n_{vu'} + \mu \frac{1}{U}}{n_{v\cdot} + \mu}$$

$$p(v'_{uo} = v | r_{uo} = u', \text{rest}) \propto \frac{n_{uv} + \eta \frac{1}{V}}{n_{u\cdot} + \eta} \cdot \frac{\sum_{u''=1}^{U} \mathbb{A}_{u''u'} n_{vu''} + \mu \frac{1}{U}}{\sum_{u''=1}^{U} \mathbb{A}_{u''\cdot} n_{vu''} + \mu}$$





Experimental Setup: Datasets & Evaluated Models

- **Twitter** datasets from [Brigadir+, WebSci '15] on the **2014 Scottish Independence Referendum** (v = Yes/No) and the **2014 US Midterm Elections** (v = Democrat/Republican)

Dataset	#Users		#Tweets	#Tokens	Vocabulary	#Interactions
	Yes/Dem.	No/Rep.				
Indyref	589	575	270,075	2,043,204	38,942	696,654
Midterms	767	778	113,545	975,199	25,312	241,741

Experimental Setup: Datasets & Evaluated Models

- **Twitter** datasets from [Brigadir+, WebSci '15] on the **2014 Scottish Independence Referendum** (v = Yes/No) and the **2014 US Midterm Elections** (v = Democrat/Republican)

Dataset	#Users		#Tweets	#Tokens	Vocabulary	#Interactions
	Yes/Dem.	No/Rep.				
Indyref	589	575	270,075	2,043,204	38,942	696,654
Midterms	767	778	113,545	975,199	25,312	241,741

- State-of-the-art baselines:
  - **Topic-Aspect Model** (TAM) from [Paul+, AAAI '10]
    $\Longrightarrow$ **Only text content** to discover viewpoints and topics
  - **Social Network Latent Dirichlet Allocation** (SN-LDA) from [Sachan+, WSDM '14]
    $\Longrightarrow$ **Text content** and **outgoing interactions** to discover **communities** ($\approx$ viewpoints) and topics
  - **Viewpoint and Opinion Discovery Unification Model** (VODUM) from [Thonet+, ECIR '16]
    $\Longrightarrow$ **Text content** to discover viewpoints and topics, and **parts of speech** to distinguish between topic words and viewpoint-topic words

Experimental Setup: Datasets & Evaluated Models

- **Twitter** datasets from [Brigadir+, WebSci '15] on the **2014 Scottish Independence Referendum** (v = Yes/No) and the **2014 US Midterm Elections** (v = Democrat/Republican)

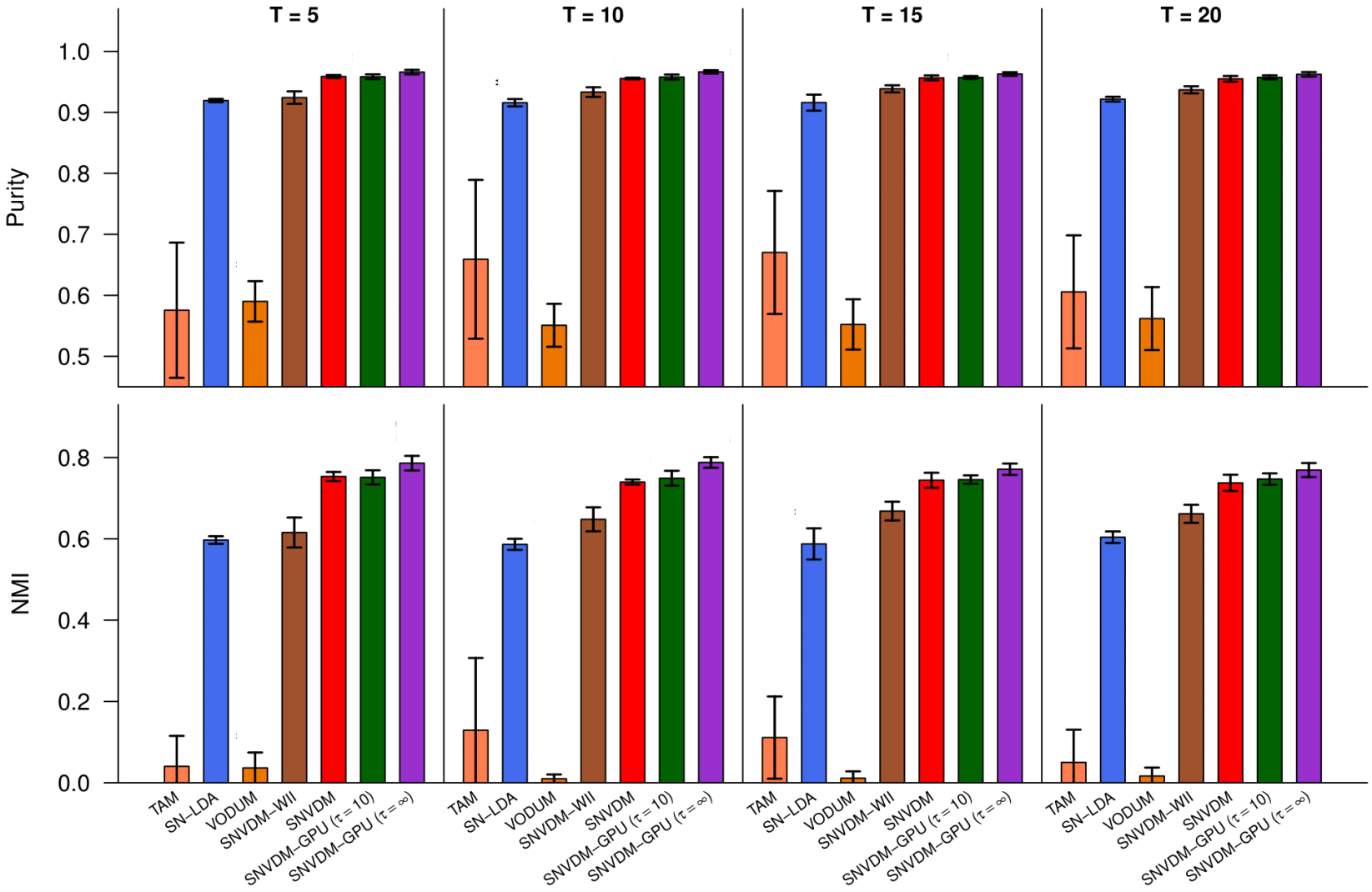
Dataset	#Users		#Tweets	#Tokens	Vocabulary	#Interactions
	Yes/Dem.	No/Rep.				
Indyref	589	575	270,075	2,043,204	38,942	696,654
Midterms	767	778	113,545	975,199	25,312	241,741

- State-of-the-art baselines:
  - **Topic-Aspect Model** (TAM) from [Paul+, AAAI '10]
  - **Social Network Latent Dirichlet Allocation** (SN-LDA) from [Sachan+, WSDM '14]
  - **Viewpoint and Opinion Discovery Unification Model** (VODUM) from [Thonet+, ECIR '16]
- Proposed models:
  - **SNVDM**
  - **SNVDM-GPU ($\tau = 10$)**: only **10 most interacting acquaintances** used in Generalized Pólya Urns
  - **SNVDM-GPU ($\tau = \infty$)**: **all acquaintances** used in Generalized Pólya Urns

Evaluation: Viewpoint Clustering

Clustering of users' viewpoints on **Indyref** in terms of **Purity** and **NMI** (error bars = 95% CI)

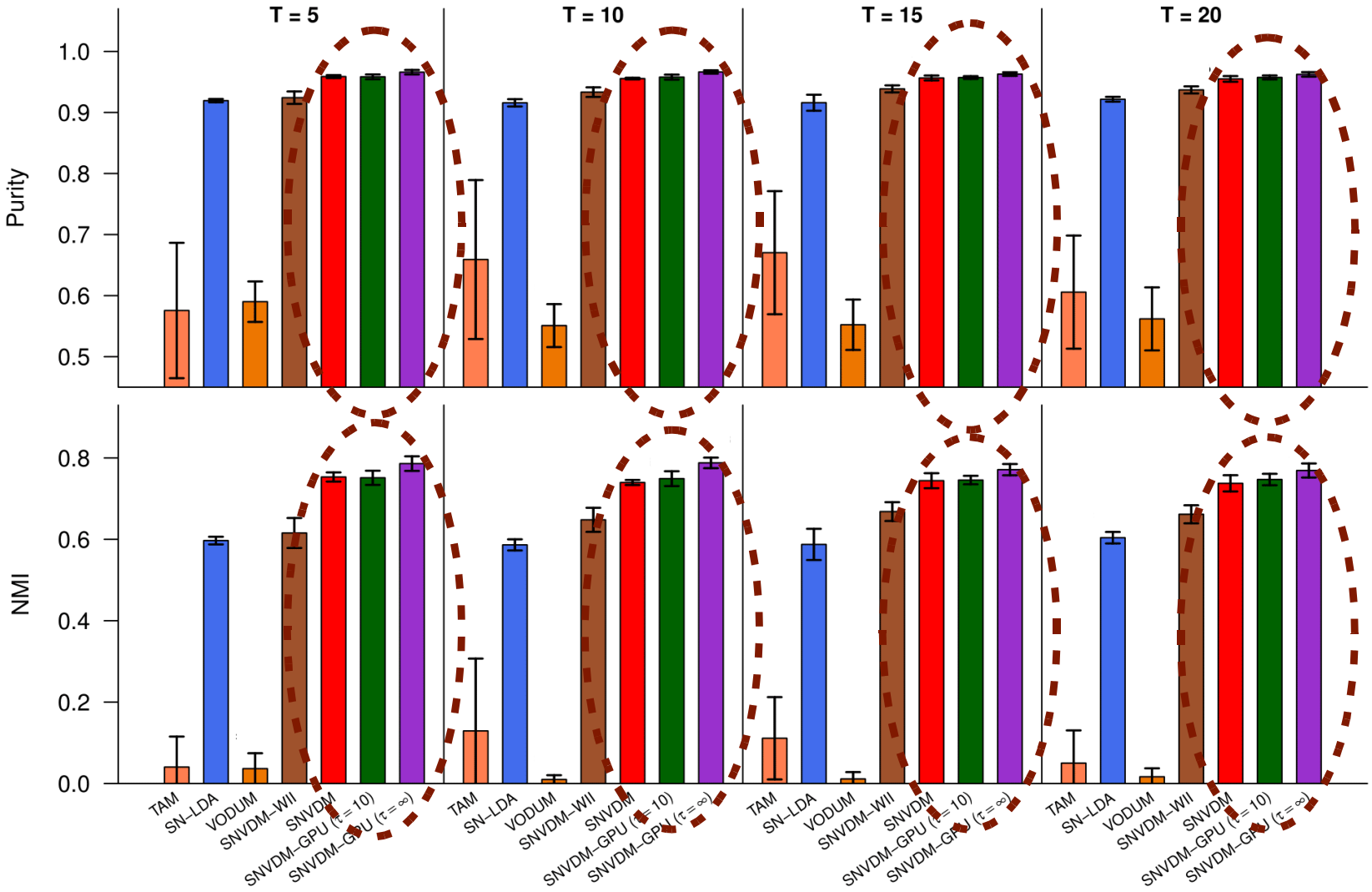
Observation 1: consistent results across different numbers of topics



Evaluation: Viewpoint Clustering

Clustering of users' viewpoints on **Indyref** in terms of **Purity** and **NMI** (error bars = 95% CI)

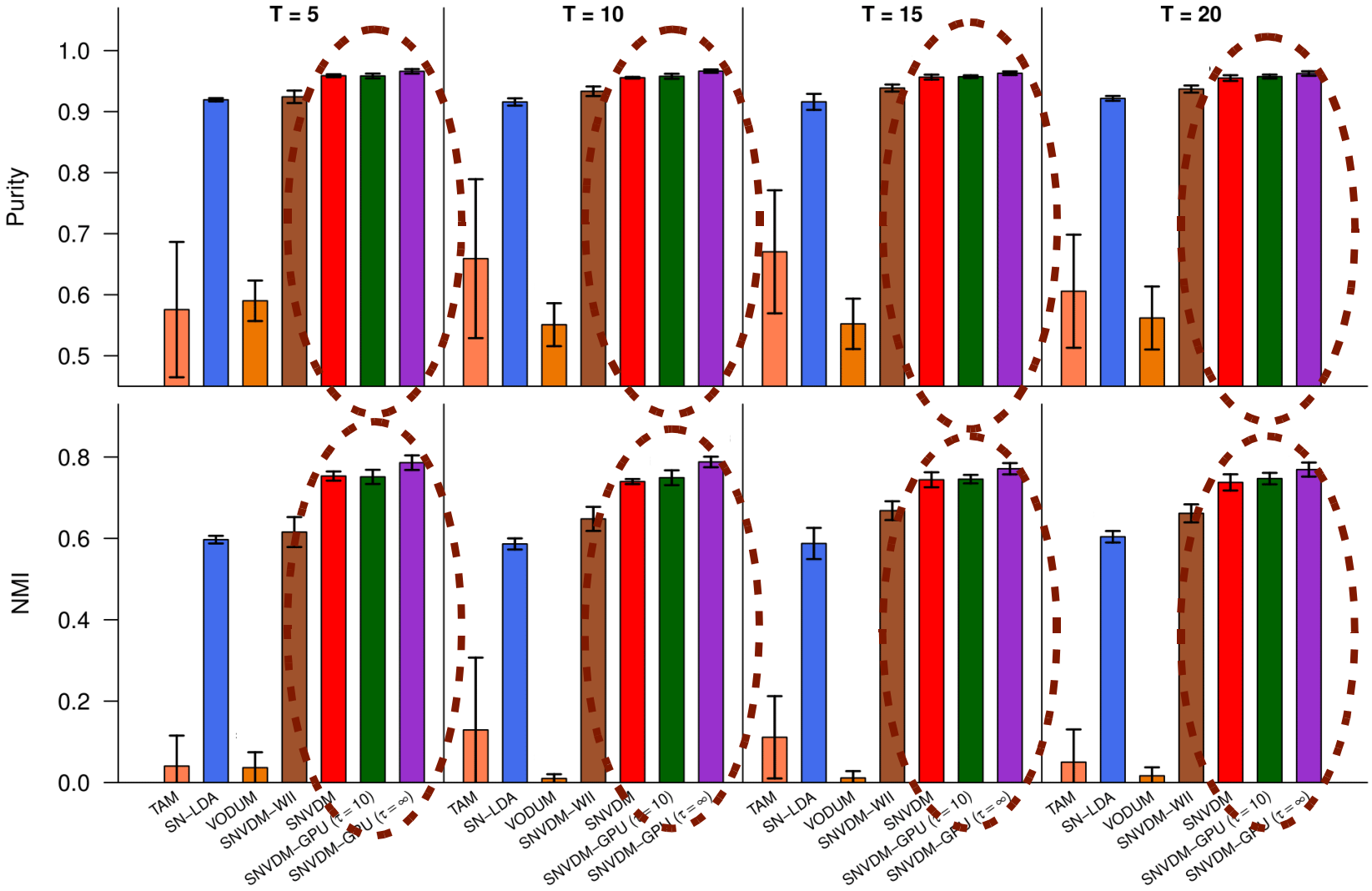
Observation 2: **SNVDM**, **SNVDM-GPU** ($\tau = 10$), **SNVDM-GPU** ($\tau = \infty$) $>$ all baselines



Evaluation: Viewpoint Clustering

Clustering of users' viewpoints on **Indyref** in terms of **Purity** and **NMI** (error bars = 95% CI)

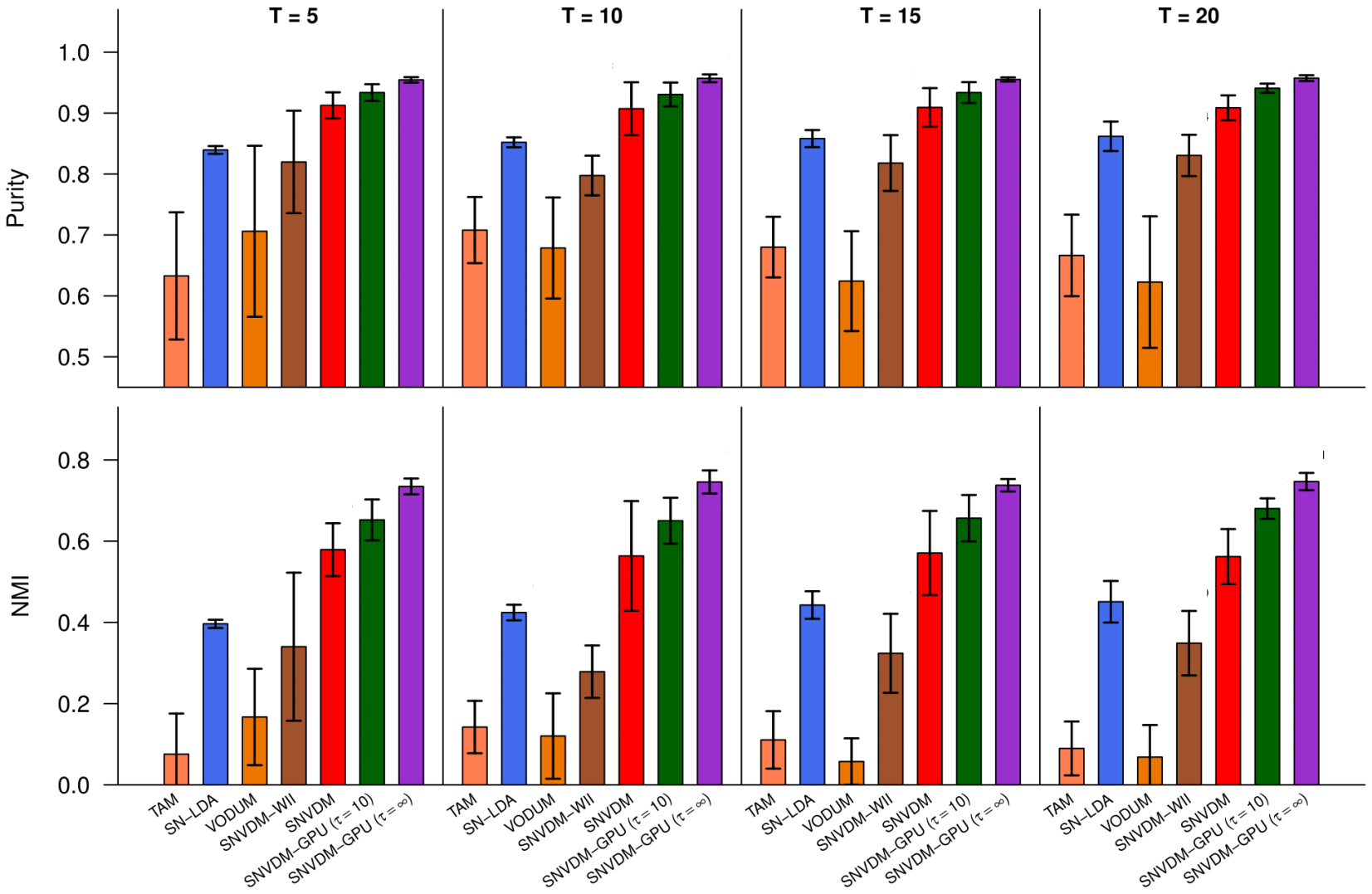
Observation 3: **SNVDM-GPU** ($\tau = \infty$) $>$ **SNVDM-GPU** ($\tau = 10$) $>$ **SNVDM** $\Longrightarrow$ GPU beneficial



Evaluation: Viewpoint Clustering

Clustering of users' viewpoints on **Midterms** in terms of **Purity** and **NMI** (error bars = 95% CI)

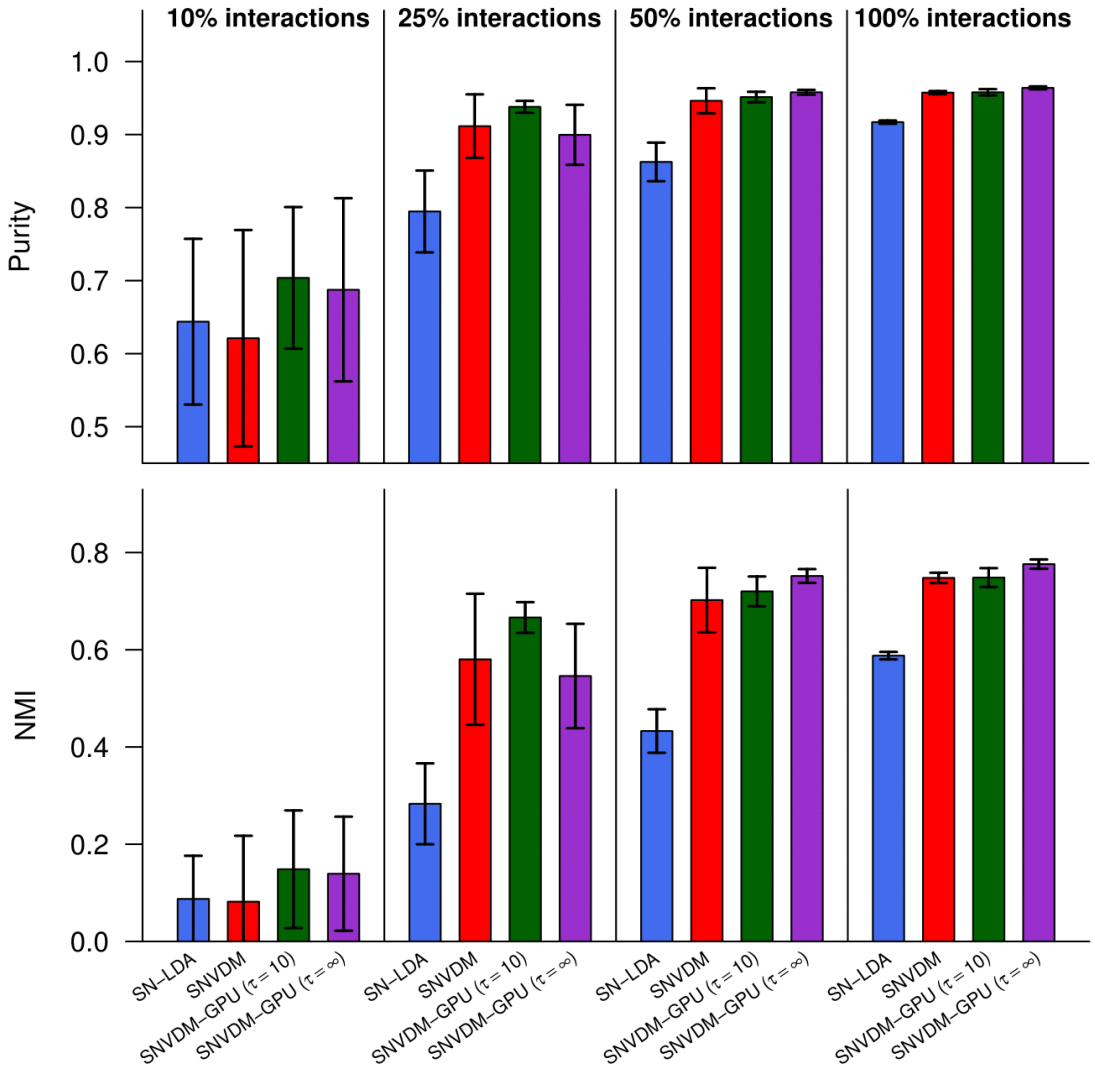
Observation 4: similar trends on **Midterms** but greater improvement for our models over baselines



Evaluation: Impact of Social Network Sparsity

Clustering of users' viewpoints on **Indyref** for different degrees of network sparsity (T = 10)

Observation: performance degraded for **lower percentage of interactions**



Evaluation: Qualitative Analysis

Most probable **topic words** and **viewpoint-topic words** for topics from Indyref and Midterms

Topic: *Scottish independence*		
Neutral	Viewpoint: *Yes*	Viewpoint: *No*
#indyref	**#voteyes**	#indyref
scotland	yes	uk
independence	scotland	salmond
vote	independence	**#bettertogether**
campaign	westminster	#scotdecides
scottish	vote	separation
uk	independent	currency
people	country	thanks
future	**#yes**	today
independent	#scotland	say

Topic: *Energy and resources*		
Neutral	Viewpoint: *Dem.*	Viewpoint: *Rep.*
energy	**#actonclimate**	**#4jobs**
house	climate	#obamacare
new	#p2	#jobs
gas	change	gop
natural	**#climatechange**	obama
#energy	clean	**bills**
#ff	oil	jobs
#kxl	energy	**house**
support	#gop	act
economic	seec	watch

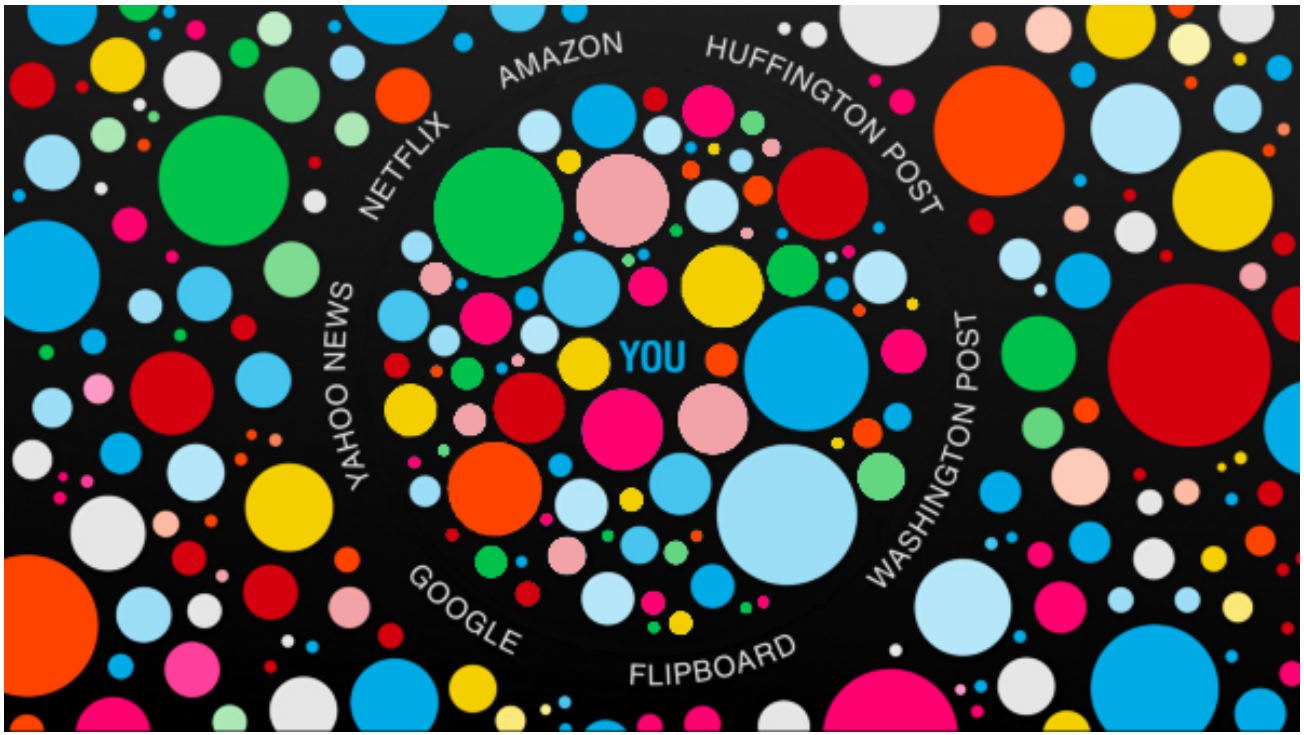
- Reasonable **coherence** of topic words and viewpoint-topic words
- Topic words indeed **unbiased** towards any viewpoints
- Use of viewpoint-specific **hashtags** and mention of different issues for different viewpoints

Conclusion and Research Directions

- **SNVDM(-GPU)**: models to jointly discover viewpoints and topics in social networks, leveraging both **posted text content** and **social interactions**

  Take-home message: **social interactions** are key for viewpoint discovery in social networks!

- What's next?
  - Integrate **time dimension** and **geolocation**, e.g., to analyze candidate support during elections
  - Design a **viewpoint summarization** framework to provide Internet users with **more diversified content** and thus mitigate the filter bubble and echo chamber phenomenon



Acknowledgments & Questions

Thanks **SIGIR**, **SIGWEB** and the **US NSF** for providing travel grants!

Questions?

@tthonet

thonet@irit.fr

@gcabanac

cabanac@irit.fr

@MohBoughanem

boughanem@irit.fr

@karenatw

sauvagnat@irit.fr

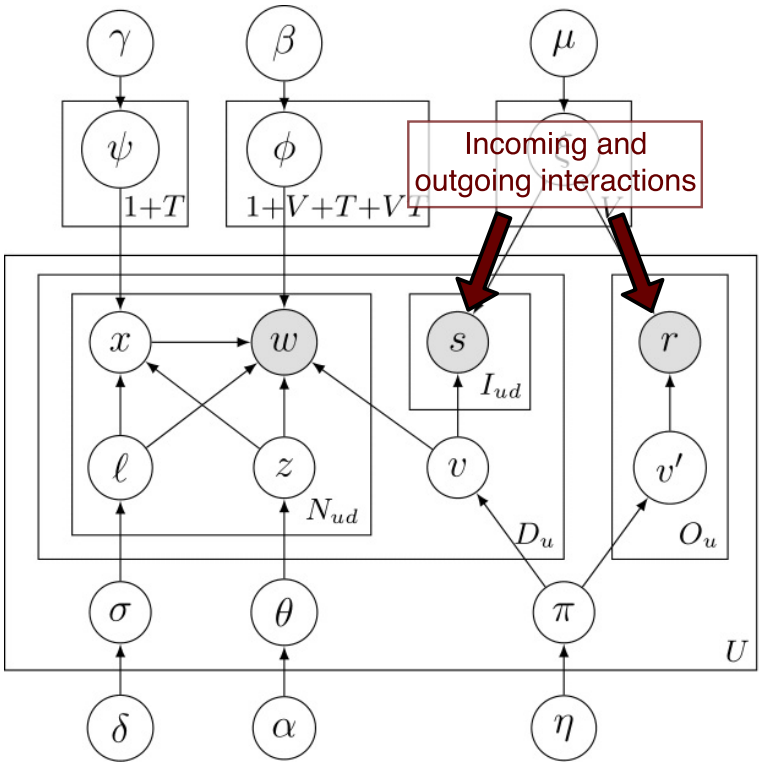
Code and data available at: `https://github.com/tthonet/SNVDM`

References

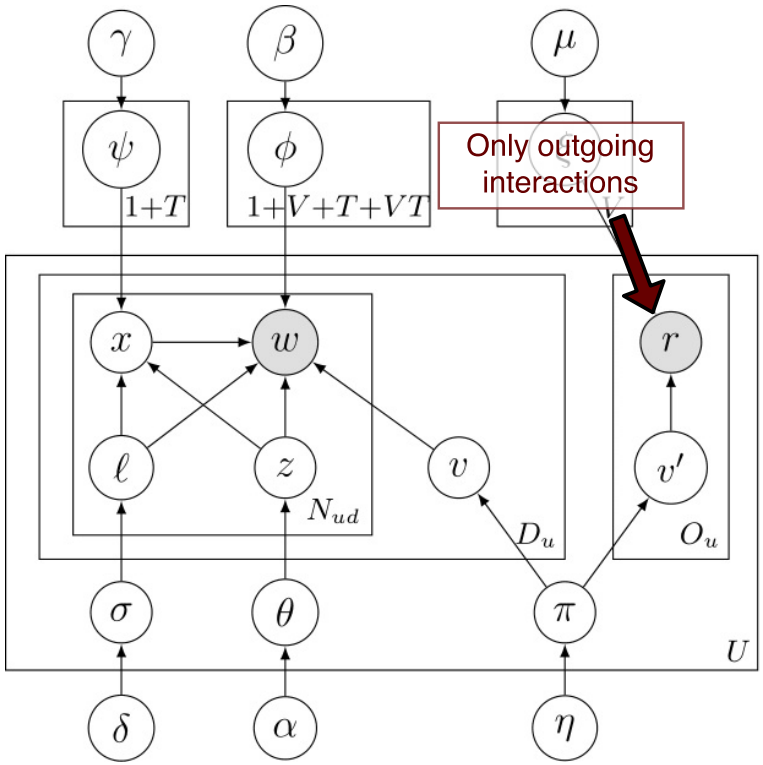
- Brigadir, I., Greene, D., & Cunningham, P. (2015). Analyzing Discourse Communities with Distributional Semantic Models. In Proc. of WebSci ’15.
- Newman, D., Asuncion, A., Smyth, P., & Welling, M. (2009). Distributed Algorithms for Topic Models. J. Mach. Learn. Res., 10, 1801–1828.
- Pariser, E. (2011). The Filter Bubble: What the Internet Is Hiding from You. The Penguin Press.
- Paul, M. J., & Girju, R. (2010). A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics. In Proc. of AAAI ’10 (pp. 545–550).
- Sachan, M., Dubey, A., Srivastava, S., Xing, E. P., & Hovy, E. (2014). Spatial Compactness meets Topical Consistency: Jointly Modeling Links and Content for Community Detection. In Proc. of WSDM ’14 (pp. 503–512).
- Sunstein, C. R. (2009). Republic.com 2.0. Princeton University Press.
- Thonet, T., Cabanac, G., Boughanem, M., & Pinel-Sauvagnat, K. (2016). VODUM: A Topic Model Unifying Viewpoint, Topic and Opinion Discovery. In Proc. of ECIR ’16 (pp. 533–545).

Appendix: Baseline SNVDM-WII

Ablated version of SNVDM: **SNVDM-WII** (without incoming interactions)



SNVDM vs . . .



. . . **SNVDM-WII**

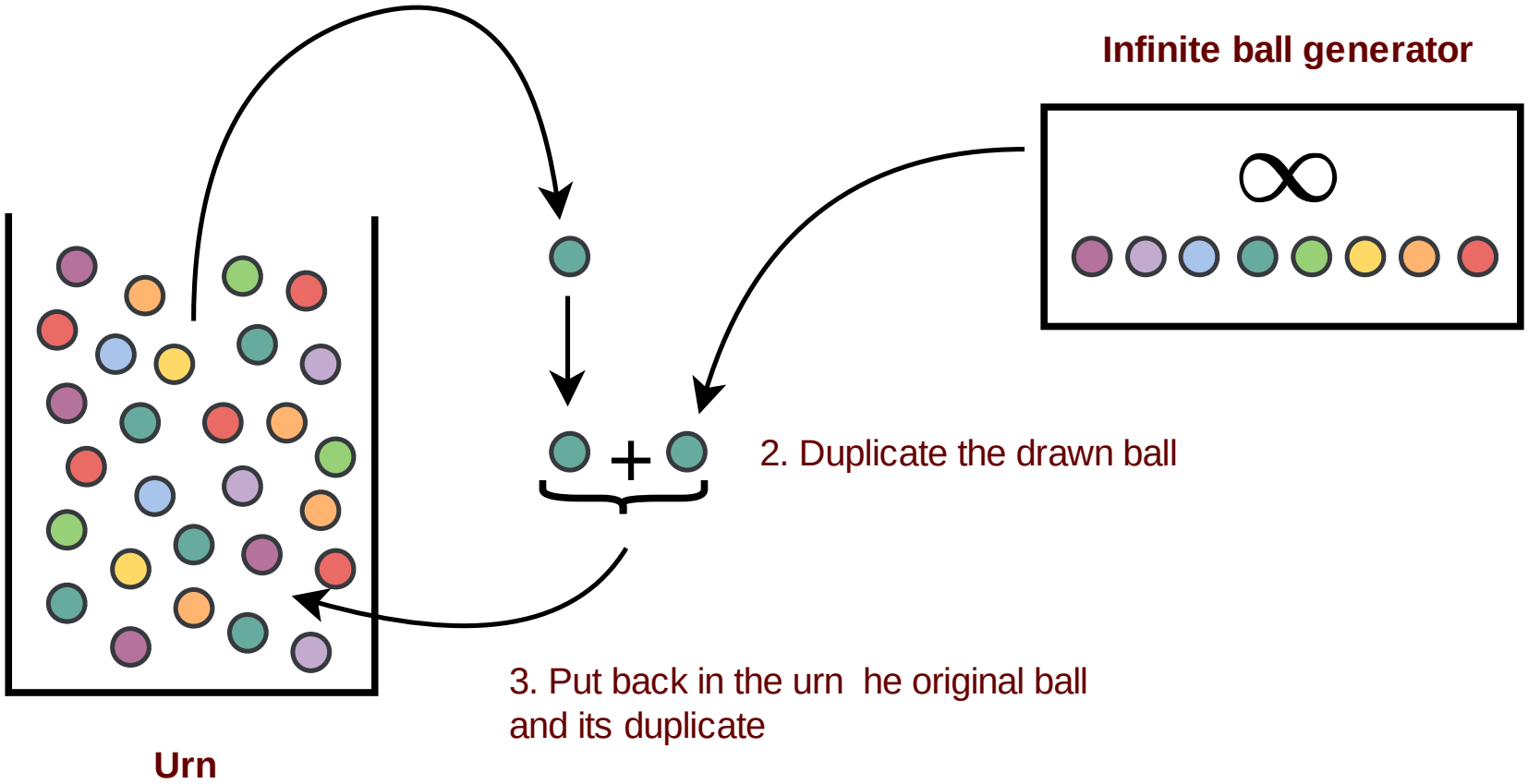
Appendix: Execution Time

Execution time (in seconds) of one Gibbs sampling iteration on Indyref (with $T = 10$) and Midterms (with $T = 15$)

	Indyref	Midterms
TAM	1.45	0.87
SN-LDA	1.18	0.64
VODUM	2.78	1.85
SNVDM-WII	2.08	1.08
SNVDM	2.49	1.15
SNVDM-GPU ($\tau = 10$)	3.47	1.34
SNVDM-GPU ($\tau = \infty$)	14.67	2.56

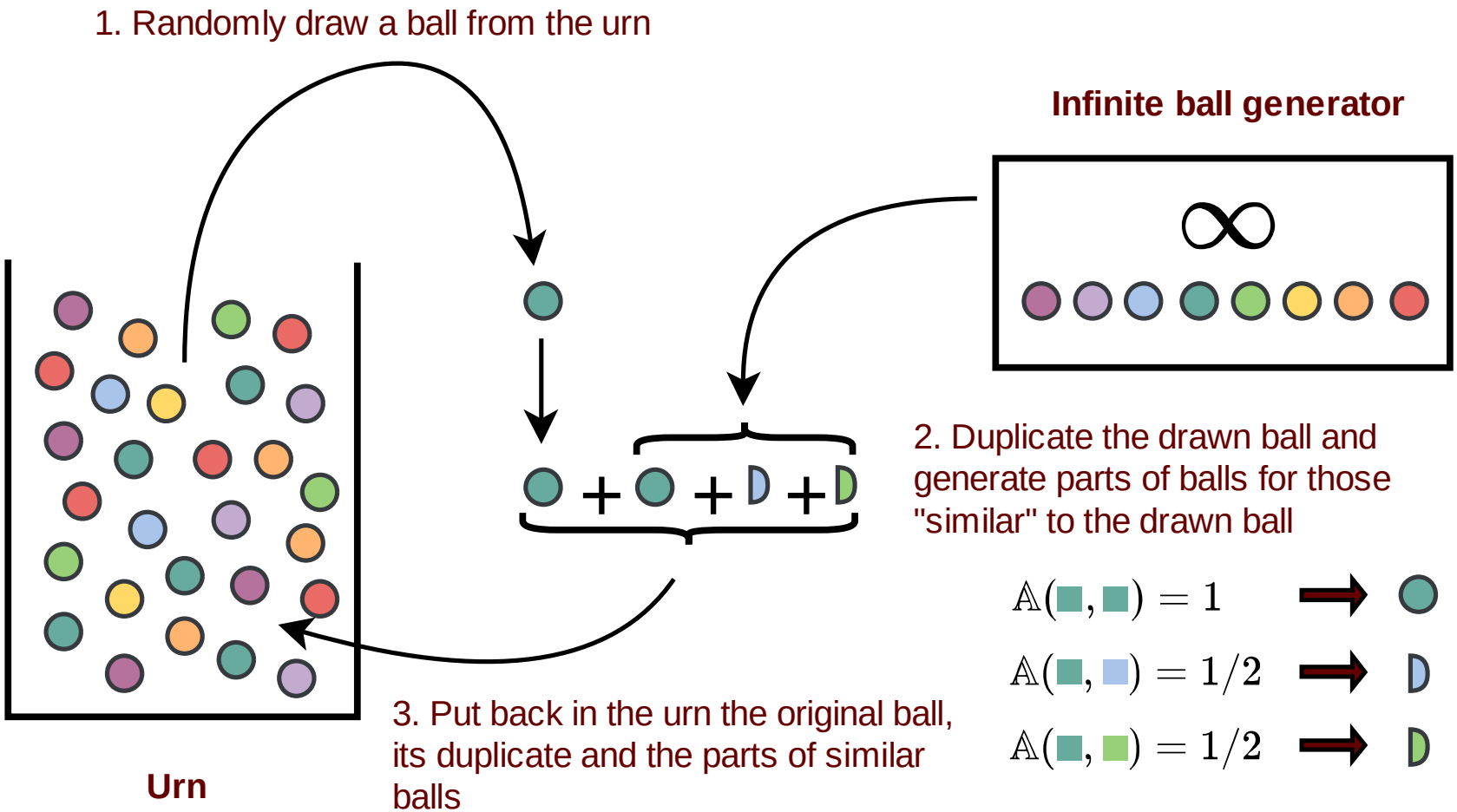
Appendix: Simple Pólya Urn Scheme

The compound **Dirichlet-Multinomial** distribution (used in LDA-based topic models) can be interpreted as an **urn sampling metaphor** with an **over-replacement** policy



Appendix: Generalized Pólya Urn Scheme

The Simple Pólya Urn scheme can be generalized by modifying the replacement rule to **exploit similarities between balls' colors** [Mahmoud, 2008]



Appendix: SNVDM-GPU

Using **Generalized Pólya Urn** in SNVDM requires minor changes in collapsed Gibbs sampling

E.g., for outgoing interaction o from user u on user u':

$$p(v'_{uo} = v | r_{uo} = u', \text{rest}) \propto \frac{n_{uv} + \eta \frac{1}{V}}{n_{u\cdot} + \eta} \cdot \frac{\boldsymbol{n_{vu'}} + \mu \frac{1}{U}}{\boldsymbol{n_{v\cdot}} + \mu}$$

SNVDM vs . . .

$$p(v'_{uo} = v | r_{uo} = u', \text{rest}) \propto \frac{n_{uv} + \eta \frac{1}{V}}{n_{u\cdot} + \eta} \cdot \frac{\sum_{u''=1}^{U} \mathbb{A}_{u''u'} \boldsymbol{n_{vu''}} + \mu \frac{1}{U}}{\sum_{u''=1}^{U} \mathbb{A}_{u''\cdot} \boldsymbol{n_{vu''}} + \mu}$$

. . . **SNVDM-GPU**

The **addition matrix** $\mathbb{A}$ defines the **weight** to put on count $n_{vu''}$ for each u'':

$$\mathbb{A}_{u'u''} = \begin{cases} 1 & \text{if } u' = u'', \\ \lambda & \text{if } u'' \text{ is among } \textbf{top } \boldsymbol{\tau} \textbf{ acquaintances of } u', \\ 0 & \text{otherwise} \end{cases}$$

with $0 \leq \lambda \leq 1$ ($\lambda = 0 \implies$ "vanilla" SNVDM) and $\tau \in \mathbb{N}$