



HAL
open science

Towards Explainability for Interaction Network Analysis

Maria Malek

► **To cite this version:**

Maria Malek. Towards Explainability for Interaction Network Analysis. ACONTA '22: First European Conference on Augmented Complex Networks - Trustworthy Analysis, Université Sorbonne Paris Nord, Nov 2022, Paris, France. hal-03986257

HAL Id: hal-03986257

<https://hal.science/hal-03986257>

Submitted on 13 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Towards Explainability for Interaction Network Analysis

Maria Malek

Aconta'22



Équipes Traitement
de l'Information
et Systèmes



Part 1

Explainability & Network analysis

- **Explainable AI (XIA)**
- **Explainability & Network Analysis**
 - Leveraging nodal and topological information
- **Network Embedding**
- **Graph Neural Network (GNN)**
- **Explainability for GNN**

Part 2

Sentiment Analysis Application

- **Explainable sentiment analysis**
- **Our approach for Detecting opinion change**
 - **Form ego Networks to Propagation network**
 - **Dataset: topological and textual information**
 - **Machine Learning and explainability**

Part 1.

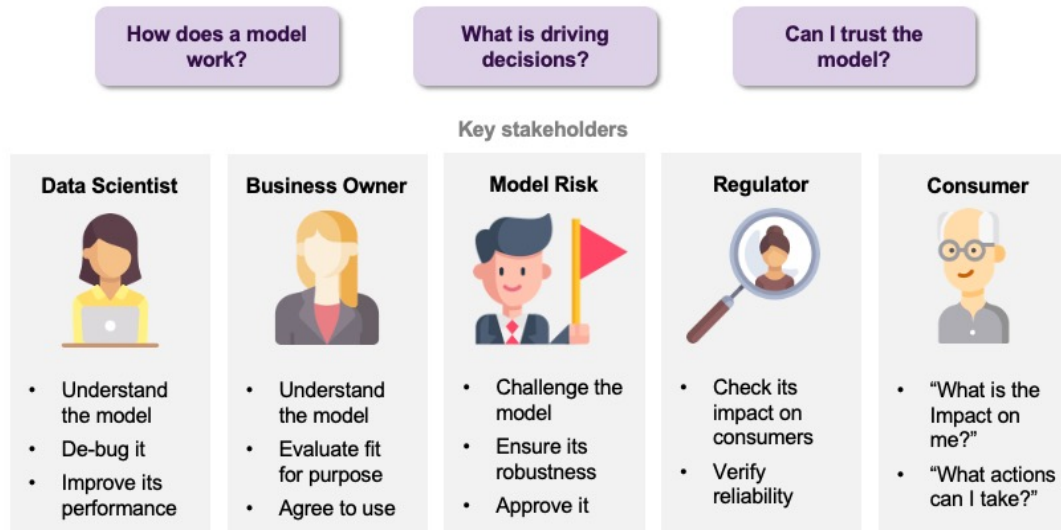
Explainability &
Network analysis

Part 1.

Explainability & Network analysis

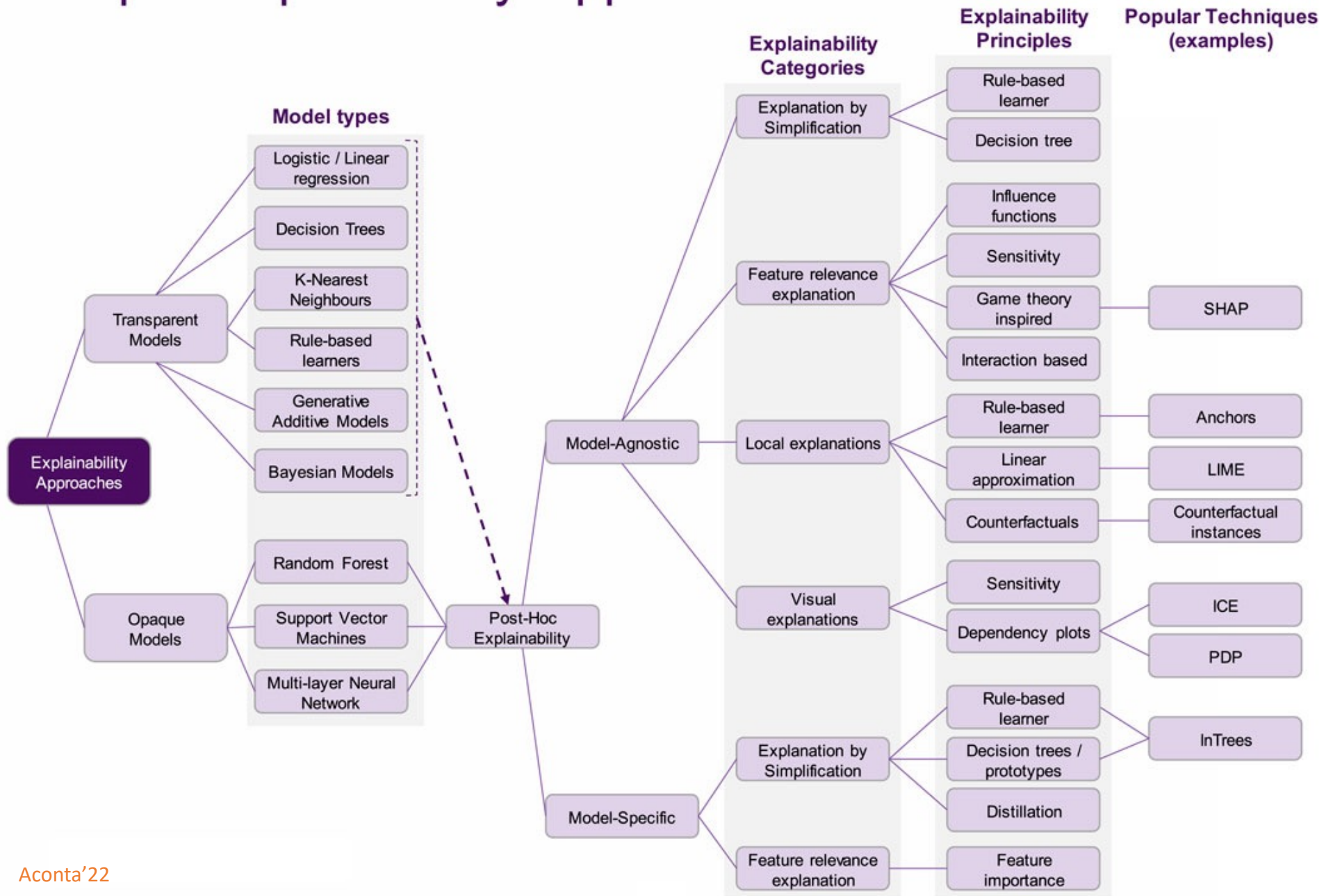
- **Explainable AI (XIA)**
- **Explainability & Network Analysis**
 - Leveraging nodal and topological information
- **Graph Embedding**
- **Graph Neural Network (GNN)**
- **Explainability for GNN**

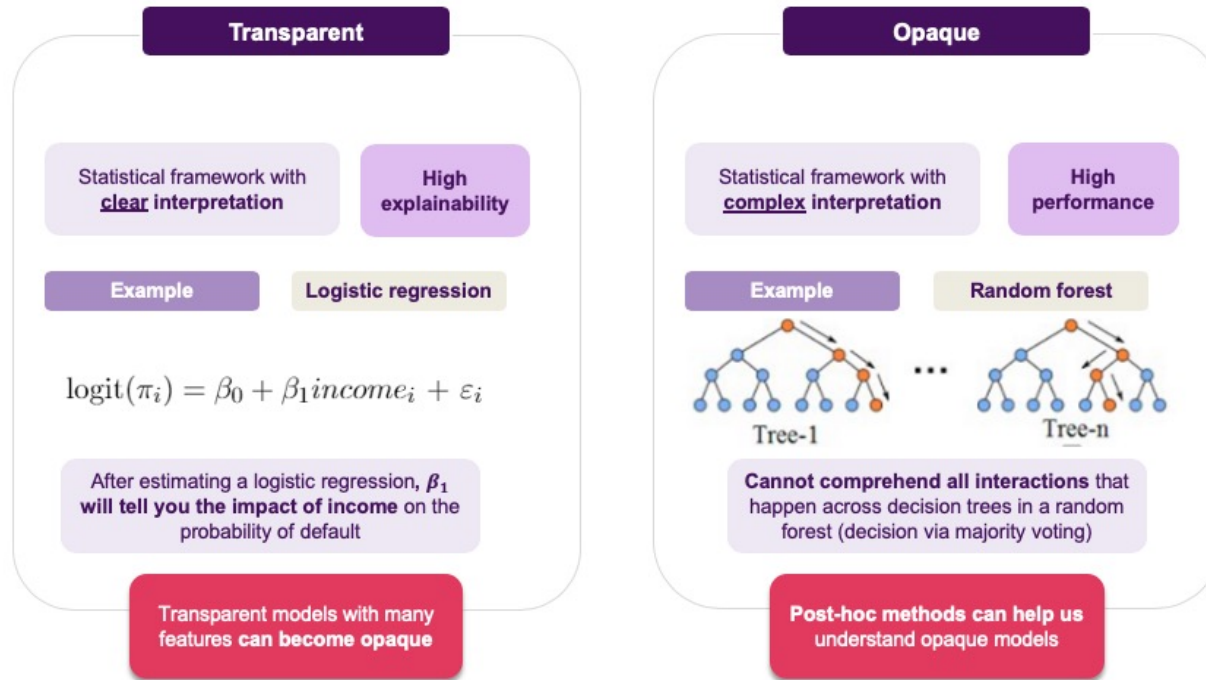
- Machine Learning methods reach today high level of performance to solve increasingly complex task.
- An emerging need to understand how decisions are rendered by AI methods
 - when the decisions of such systems affect the lives of human.
- The paradigm underlying this problem is the so-called Explainable AI (XIA) domain



[V Belle, I Papantonis , Principles and practice of explainable machine learning, - Frontiers in big Data, 2021](#)

Map of Explainability Approaches





[V Belle, I Papantonis , Principles and practice of explainable machine learning, - Frontiers in big Data, 2021](#)

| Explanation | Advantages | Disadvantages |
|--------------------|--|--|
| Local explanations | Explains the model's behaviour in a local area of interest. Operates on instance-level explanations. | Explanations do not generalize on a global scale. Small perturbations might result in very different explanations. Not easy to define locality. Some approaches face stability issues. |
| Examples | Representative examples provide insights about the model's internal reasoning. Some of the algorithms uncover the most influential training data points that led the model to its predictions. | Examples require human inspection. They do not explicitly state what parts of the example influence the model. |
| Feature relevance | They operate on an instance level, calculating the importance of each feature in the model's decision. A number of the proposed approaches come with appealing theoretical guarantees. | They are sensitive in cases where the features are highly correlated. In many cases the exact solutions are approximated, leading to undesirable side effects, such as the ordering affecting the outcome. |
| Simplification | Simple surrogate models explain the opaque ones. Resulting explanations, such as rules, are easy to understand. | Surrogate models may not approximate the original models well. Surrogate models come with their own limitations. |
| Visualizations | Easier to communicate to non technical audience. Most of the approaches are intuitive and not hard to implement. | There is an upper bound on how many features we can consider at once. Humans need to inspect the resulting plots in order to produce explanations. |

[V Belle, I Papantonis , Principles and practice of explainable machine learning, - Frontiers in big Data, 2021](#)

Part 1.

Explainability &
Network analysis

Part 1.

Explainability & Network analysis

- **Explainable AI (XIA)**
- **Explainability & Network Analysis**
 - Leveraging nodal and topological information
- **Network Embedding**
- **Graph Neural Network (GNN)**
- **Explainability for GNN**

- Extend the explainability approaches to the domain of complex network analysis
 - produce emergent explanations from topological information:
 - global topological measures (density, clustering coefficient, diameter)
 - local measures as the different types of centralities: degree, betweenness, pageRank, etc.
- Necessary when complex network analysis is part of a decision system to perform a task:
 - recommender systems, sentiment analysis or link prediction.
- Useful for understanding the community structures detected in the network.

Spyroula Masiala and Martin Atzmueller (2018) First Perspectives on Explanation in Complex Network Analysis. In: Proc. Benelux Conference on Artificial Intelligence (BNAIC), Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

- **Machine learning**: entities processed in interpretability method are related to
 - features (attributes), data, extracted prototypes
 - simple extracted knowledge often in the form of rules or decision trees.

- **Interaction networks**: new entities must be taken into consideration:
 - nature of the links in the graph (social link, collaboration or interaction link, etc.),
 - the topological information extracted from the graph.
 - **To propose complete explanatory actions** which integrate topological and semantic information.

Idea 1: Apply explainable ML to interaction networks

Network embedding and explainable GNN

Idea 2: Define an explainability approaches for graph (network) algorithms

Part 1.

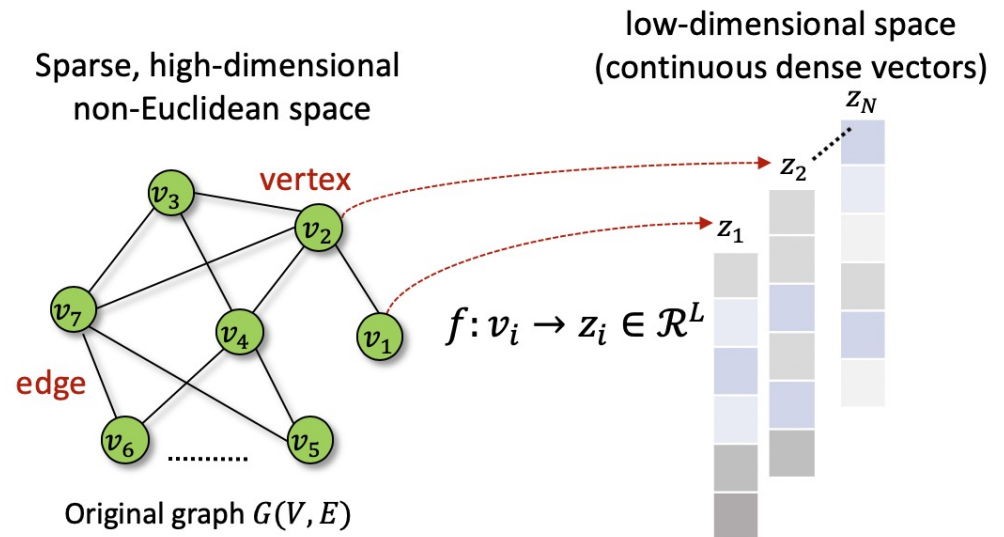
Explainability &
Network analysis

Part 1.

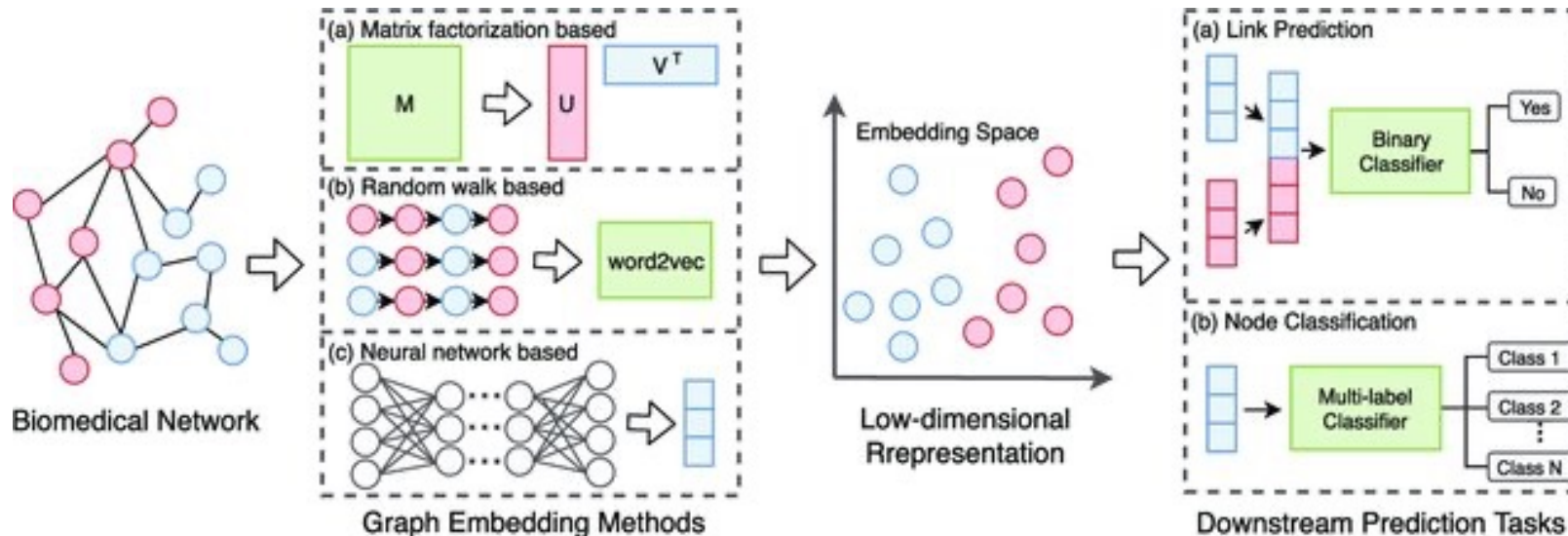
Explainability & Network analysis

- **Explainable AI (XIA)**
- **Explainability & Network Analysis**
 - Leveraging nodal and topological information
- **Network Embedding**
- **Graph Neural Network (GNN)**
- **Explainability for GNN**

- Industrial systems: graphs with over 50 million nodes and over a billion edges
 - Need of low computational complexity algorithms.
- Network embedding: learning latent low-dimensional feature representations for the nodes or links in a network.
- Learn encodings for the nodes in the network such that
 - the similarity in the embedding space reflects the similarity in the network.

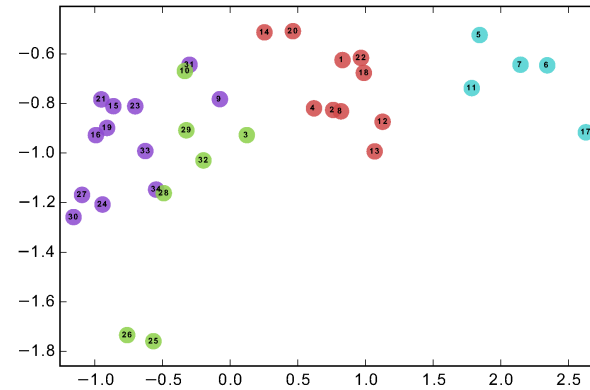
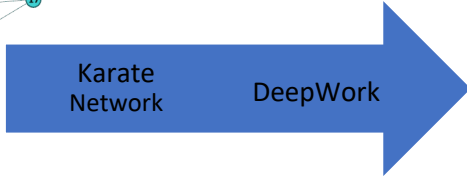
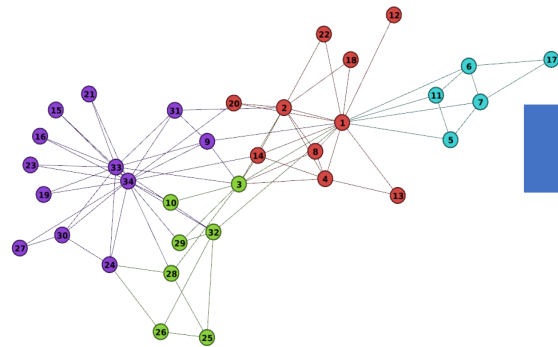


Xu, Mengjia, [Understanding Graph Embedding Methods and Their Applications](#), SIAM Review, Vol. 63, No.44, pp. 825-853, 2021

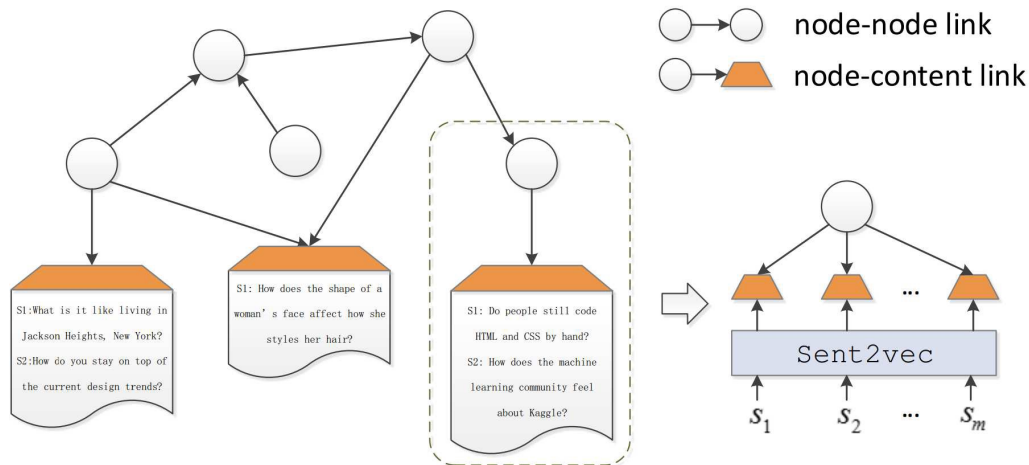


Xiang Yue, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yungui Huang, Simon M Lin, Wen Zhang, Ping Zhang, Huan Sun, Graph embedding on biomedical networks: methods, applications and evaluations, *Bioinformatics*, Volume 36, Issue 4, 15 February 2020, Pages 1241–1251, <https://doi.org/10.1093/bioinformatics/btz718>

- Support noisy datasets across different application domains:
 - social networks, citation networks, language networks, and biological networks



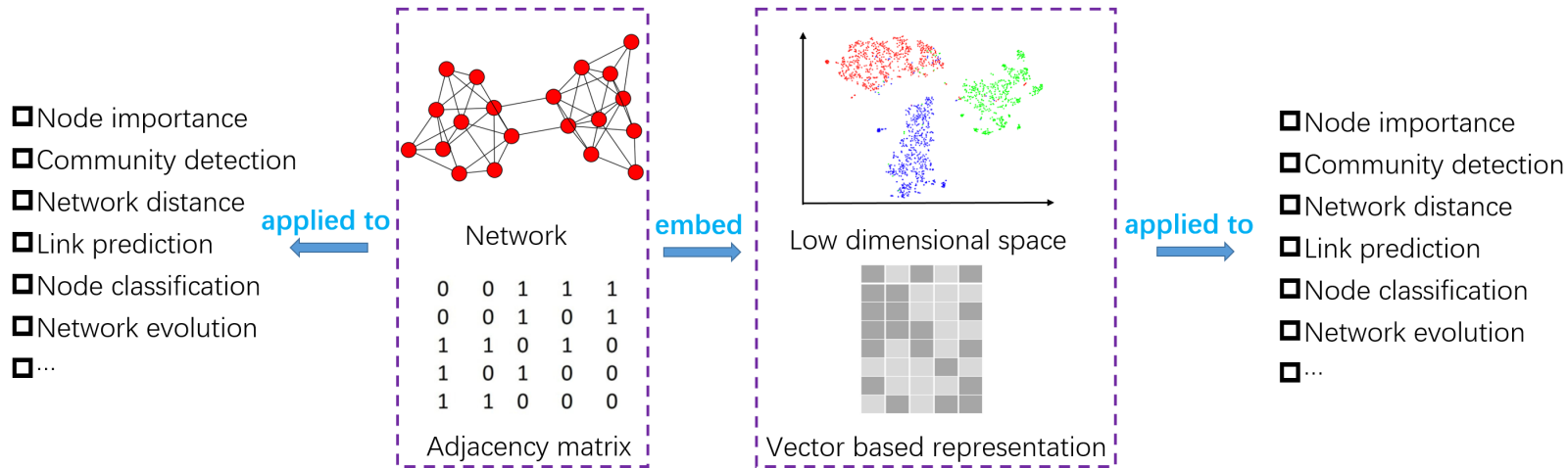
Bryan Perozzi and Rami Al-Rfou and Steven Skiena, DeepWalk, Proceedings of the 20th ACM, SIGKDD, international conference on Knowledge discovery and data mining, 2014



Sun, X.; Guo, J.; Ding, X.; and Liu, T., A general framework for content-enhanced network representation learning. arXiv preprint arXiv:1610.02906, 2016

Traditional topology based network analysis

Network embedding based network analysis



P. Cui, X. Wang, J. Pei and W. Zhu, "A Survey on Network Embedding,
IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 5, pp. 833-852, 1 May 2019,

Despite the superior performances, one fundamental limitation of them is the lack of interpretability (Liu N. et al., 2018). Different dimensions in the embedding space usually have no specific meaning, thus it is difficult to comprehend the underlying factors that have been preserved in the latent space.

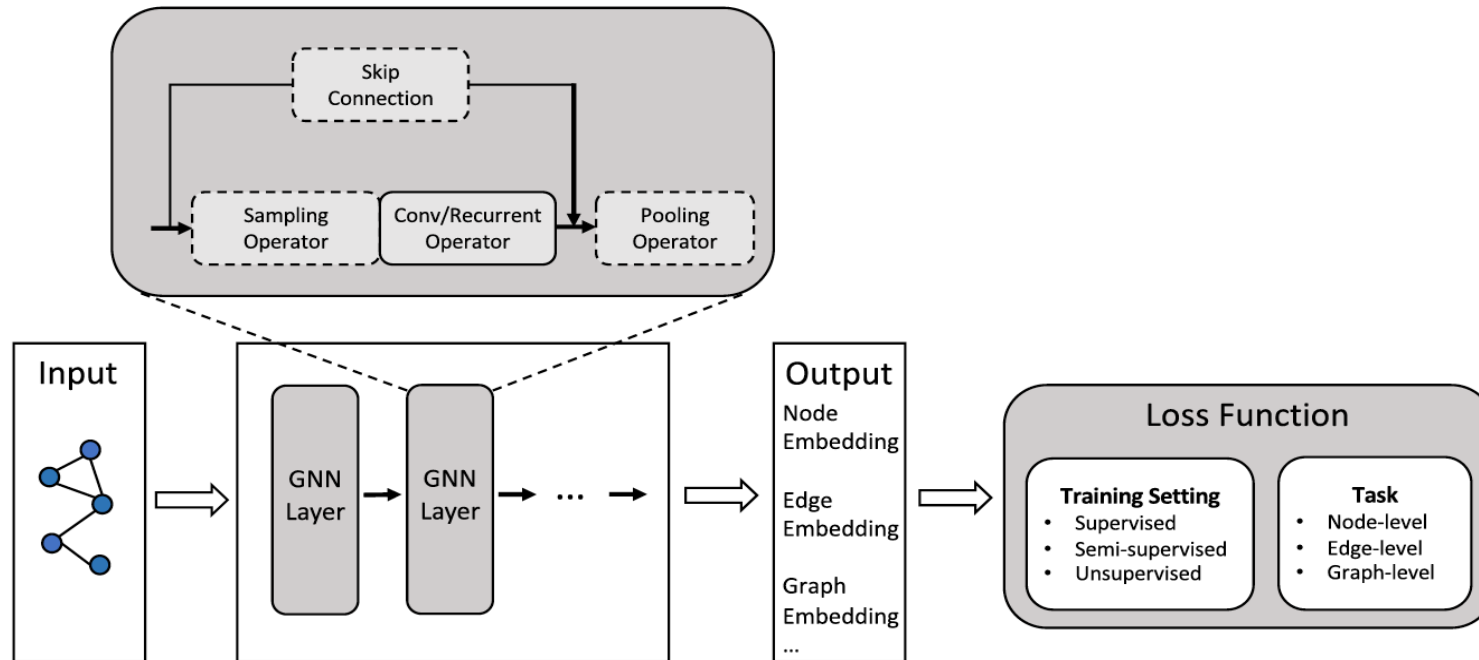
Part 1.

Explainability &
Network analysis

Part 1.

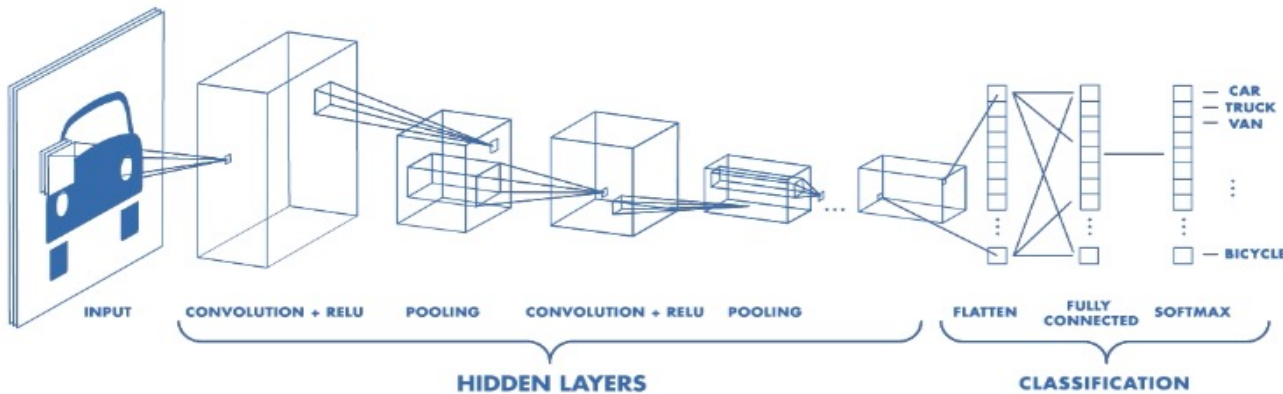
Explainability & Network analysis

- **Explainable AI (XIA)**
- **Explainability & Network Analysis**
 - Leveraging nodal and topological information
- **Network Embedding**
- **Graph Neural Network (GNN)**
- **Explainability for GNN**

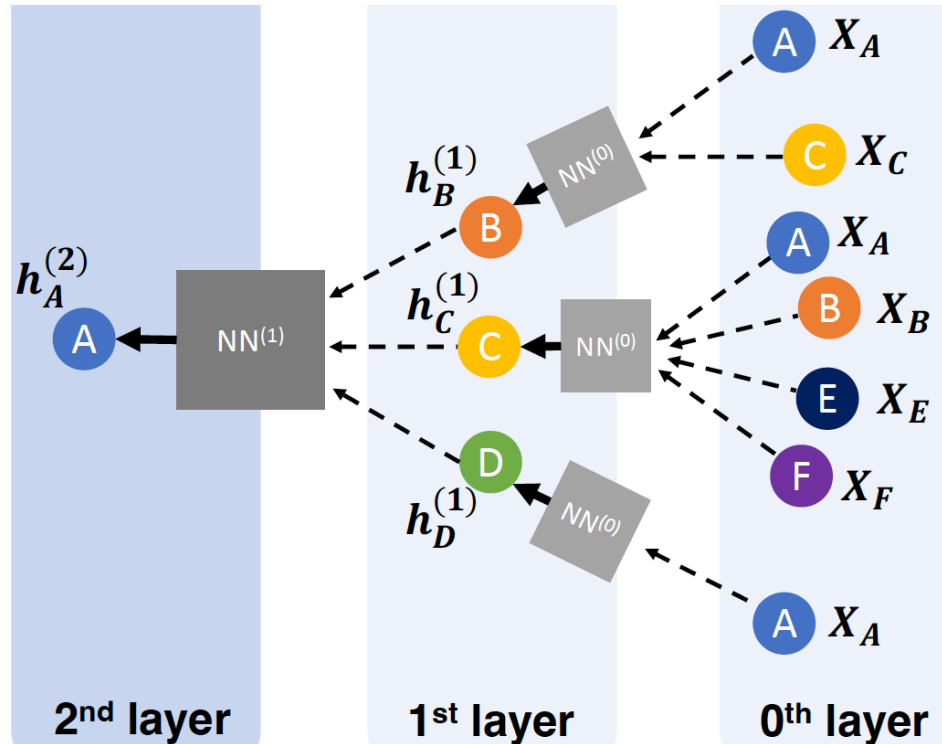
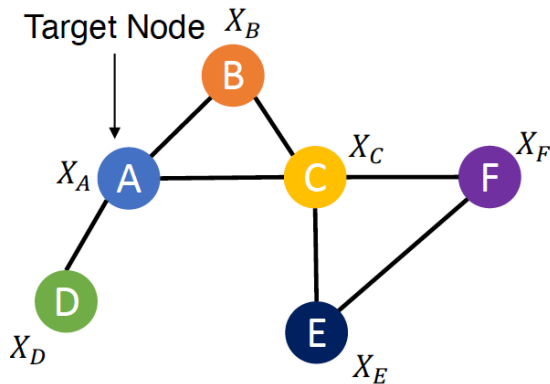


- Graph neural networks (GNNs) are deep learning-based methods that operate on graph domain
- Applications: Citation networks, Bio-chemical Graphs, Social networks, Knowledge graphs

[Jie Zhou](#), [Ganqu Cui](#), [Shengding Hu](#), [Zhengyan Zhang](#), [Cheng Yang](#), [Zhiyuan Liu](#), [Lifeng Wang](#), [Changcheng Li](#), [Maosong Sun](#), **Graph Neural Networks: A Review of Methods and Applications**, AI Open, 2021



- Deep neural networks, especially convolutional neural networks (CNNs) (LeCun et al., 1998)
 - CNNs have the ability to extract multi-scale localized spatial features
 - compose them to construct highly expressive representations,
 - led to in almost all machine learning areas (LeCun et al., 2015).
- The keys of CNNs are local connection, shared weights and the use of multiple layers
- CNNs operate on regular Euclidean data like images (2D grids) and texts (1D sequences)



1. Aggregate messages

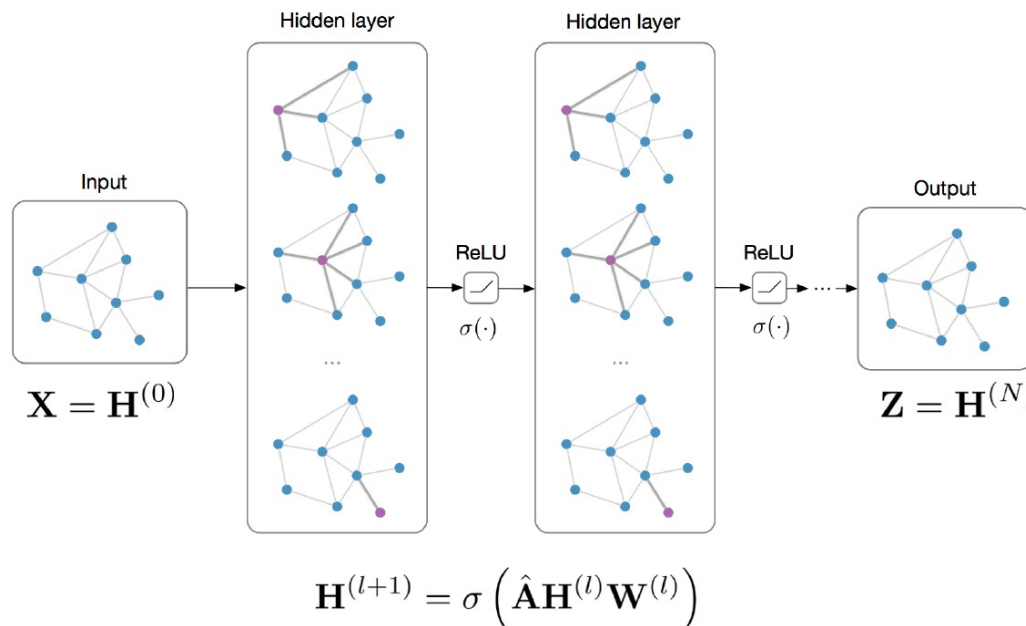
$$m_v^{(l)} = \frac{1}{|\mathcal{N}(v) + 1|} \sum_{u \in \mathcal{N}(v) \cup \{v\}} h_u^{(l)}$$

2. Transform messages

$$h_v^{(l+1)} = \mathbf{W}^{(l)} \circ m_v^{(l)}$$

Minji Yoon (CMU) - Guest lecture at 10707: Introduction to Deep Learning

Input: Feature matrix $\mathbf{X} \in \mathbb{R}^{N \times E}$, preprocessed adjacency matrix $\hat{\mathbf{A}}$



Node classification:

$$\text{softmax}(\mathbf{z}_n)$$

e.g. Kipf & Welling (ICLR 2017)

Graph classification:

$$\text{softmax}(\sum_n \mathbf{z}_n)$$

e.g. Duvenaud et al. (NIPS 2015)

Link prediction:

$$p(A_{ij}) = \sigma(\mathbf{z}_i^T \mathbf{z}_j)$$

Kipf & Welling (NIPS BDL 2016)

“Graph Auto-Encoders”

Thomas Kipf, University of Amsterdam

| Dataset | Type | Nodes | Edges | Classes | Features | Label rate |
|----------|------------------|--------|---------|---------|----------|------------|
| Citeseer | Citation network | 3,327 | 4,732 | 6 | 3,703 | 0.036 |
| Cora | Citation network | 2,708 | 5,429 | 7 | 1,433 | 0.052 |
| Pubmed | Citation network | 19,717 | 44,338 | 3 | 500 | 0.003 |
| NELL | Knowledge graph | 65,755 | 266,144 | 210 | 5,414 | 0.001 |

| Method | Citeseer | Cora | Pubmed | NELL |
|-------------------------|------------------|------------------|-------------------|-------------------|
| ManiReg [3] | 60.1 | 59.5 | 70.7 | 21.8 |
| SemiEmb [28] | 59.6 | 59.0 | 71.1 | 26.7 |
| LP [32] | 45.3 | 68.0 | 63.0 | 26.5 |
| DeepWalk [22] | 43.2 | 67.2 | 65.3 | 58.1 |
| ICA [18] | 69.1 | 75.1 | 73.9 | 23.1 |
| Planetoid* [29] | 64.7 (26s) | 75.7 (13s) | 77.2 (25s) | 61.9 (185s) |
| GCN (this paper) | 70.3 (7s) | 81.5 (4s) | 79.0 (38s) | 66.0 (48s) |
| GCN (rand. splits) | 67.9 ± 0.5 | 80.1 ± 0.5 | 78.9 ± 0.7 | 58.4 ± 1.7 |

Summary of results in terms of classification accuracy (in %)

Kipf, Thomas N., and Max Welling,
Semi-supervised classification with graph convolutional networks, arXivpreprint arXiv:1609.02907, (2016)

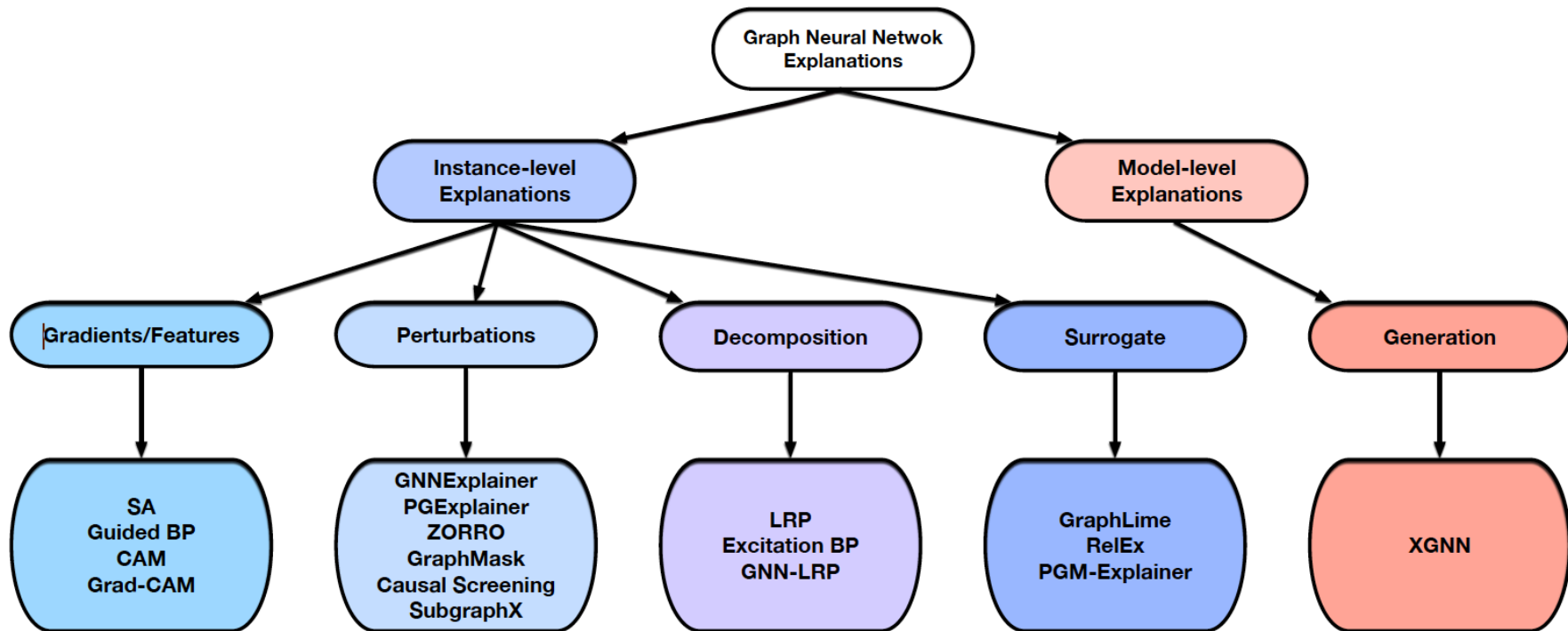
Part 1.

Explainability &
Network analysis

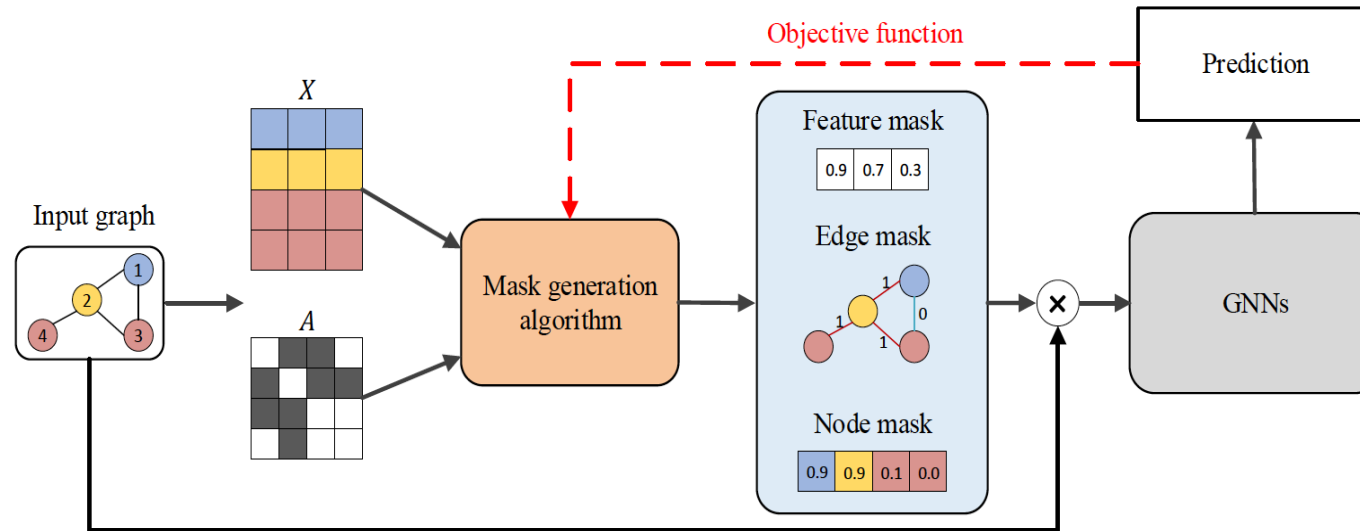
Part 1.

Explainability & Network analysis

- **Explainable AI (XIA)**
- **Explainability & Network Analysis**
 - Leveraging nodal and topological information
- **Network Embedding**
- **Graph Neural Network (GNN)**
- **Explainability in GNN**



Yuan, Hao and Yu, Haiyang and Gui, Shurui and Ji, Shuiwang, Explainability in Graph Neural Networks: A Taxonomic Survey, arXiv, 2020,



- Mask generation algorithms to obtain different types of masks.
- The mask is combined with the input graph to capture important input information.
- The trained GNNs evaluate whether the new prediction is similar to the original prediction
 - can provide guidance for improving the mask generation algorithms.

*Intuitively,
when important input information is retained,
the predictions should be similar to the original
predictions.*

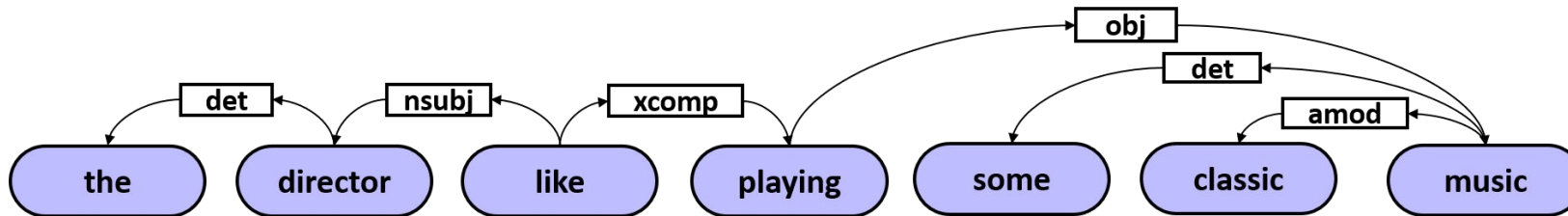
- **Fidelity+** : the difference of accuracy (or predicted probability) between the original predictions and the new predictions after masking out important input features (nodes/edges/ node feature)
- Explanations should be **sparse**, they should capture the most important input features and ignore the irrelevant ones.
- The metric **Sparsity** measures such a property.
 - measures the fraction of features selected as important by explanation methods

Accuracy related to ground truth (F1 score, AUC)
In addition, good explanations should be **stable**.

Stability: when small changes are applied to the input without affecting the predictions, the explanations should remain similar.

- 3 sentiment graph datasets based on text sentiment analysis data : SST2, SST5, and Twitter,
- Molecule data : BBBP
- Synthetic data sets: BA-shapes, BA-2motifs

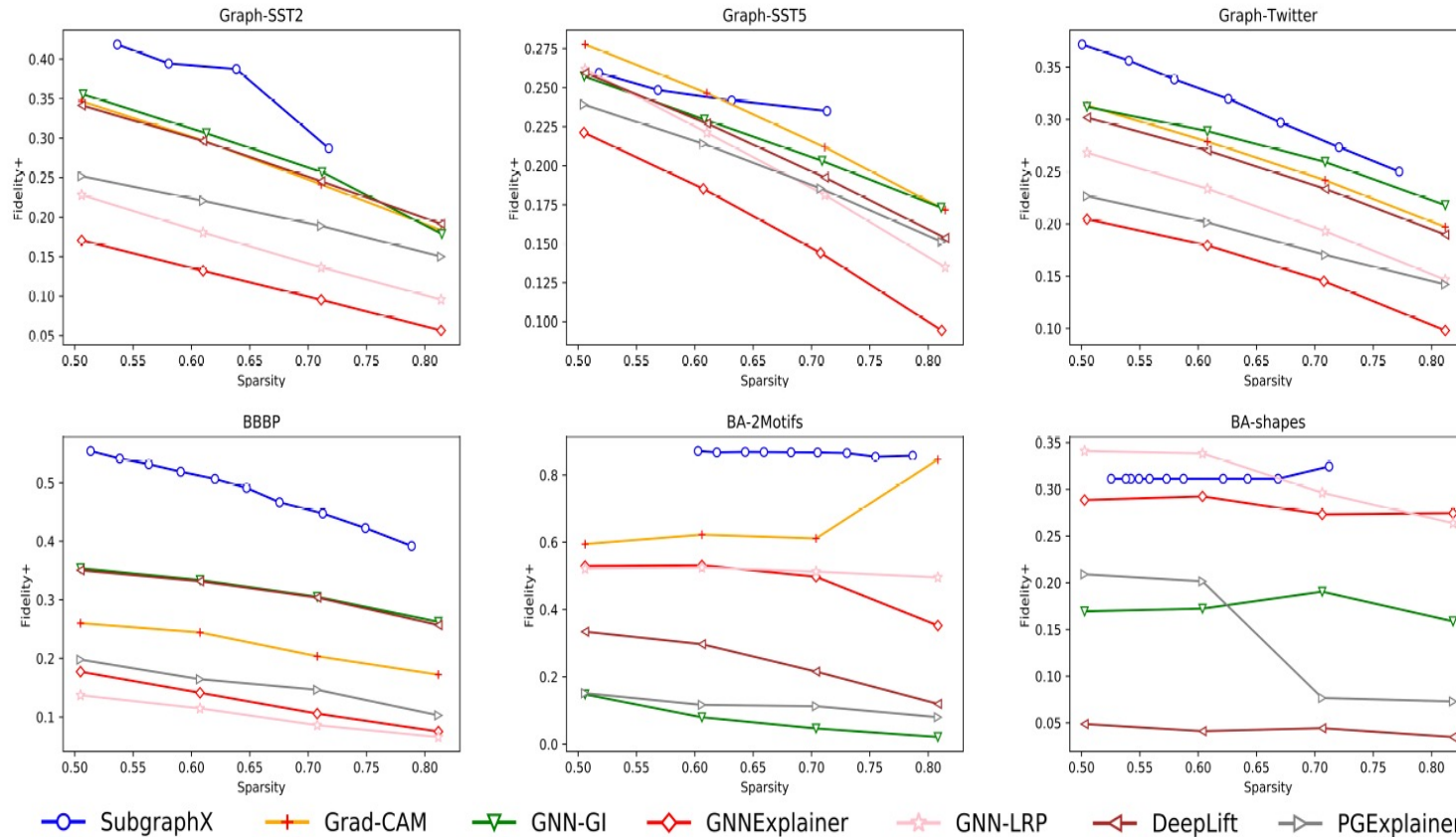
| Dataset | Graph-SST2 | Graph-SST5 | Graph-Twitter |
|-------------------|------------|------------|---------------|
| # of classes | 2 | 5 | 3 |
| # of features | 768 | 768 | 768 |
| Avg. # of nodes | 10.199 | 19.849 | 21.103 |
| # of train graphs | 67,349 | 8,544 | 4,998 |
| # of val. graphs | 872 | 1,101 | 1,250 |
| # of test graphs | 1,821 | 2,210 | 692 |
| GCN accuracy | 0.892 | 0.443 | 0.614 |



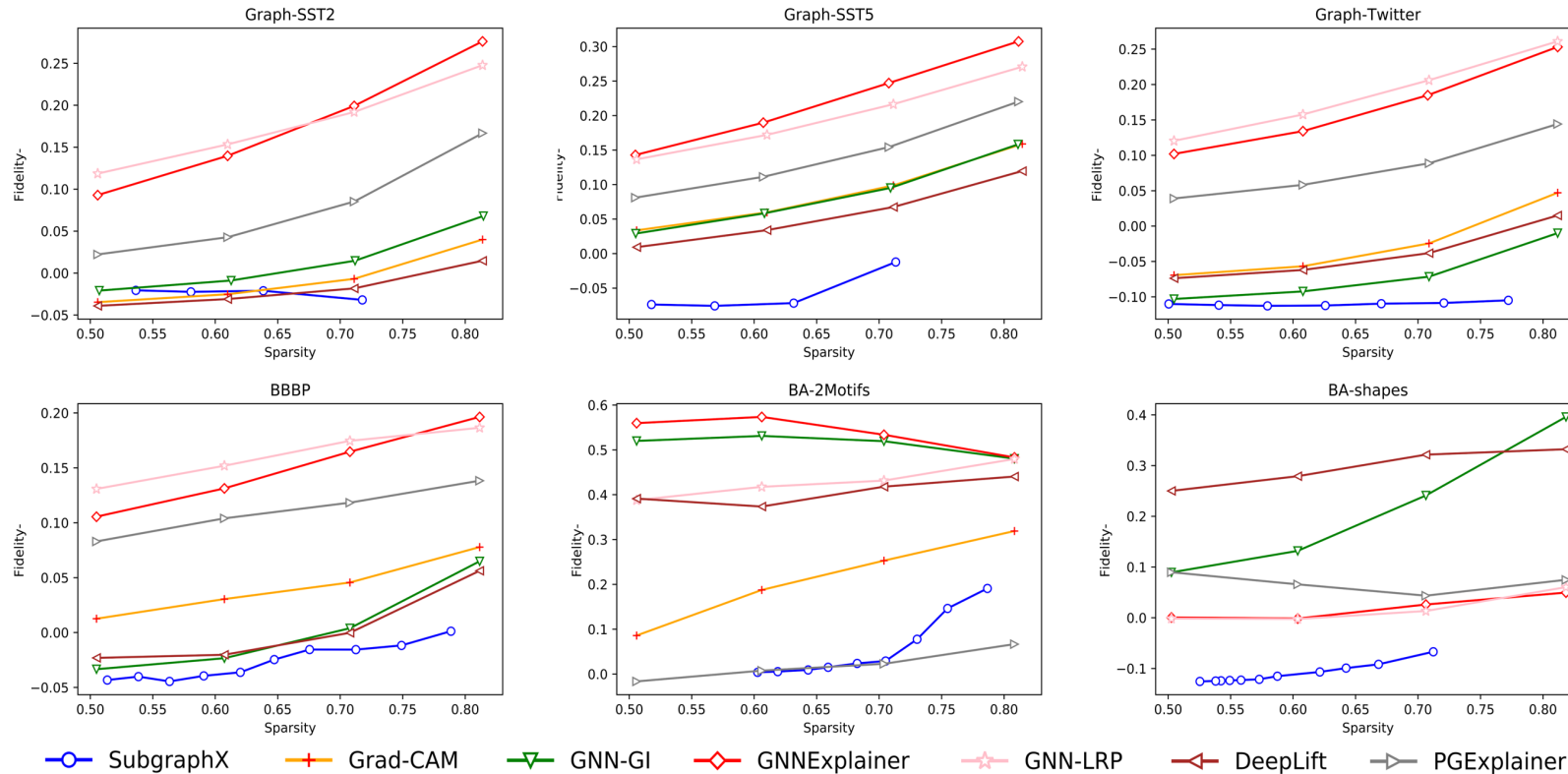
- **The Fidelity+ metric** studies the prediction change by removing important nodes/edges/node features.

- **the metric Fidelity-** studies prediction change by keeping important input features and removing unimportant features.

Yuan, Hao and Yu, Haiyang and Gui, Shurui and Ji, Shuiwang, Explainability in Graph Neural Networks: A Taxonomic Survey, arXiv, 2020,



Yuan, Hao and Yu, Haiyang and Gui, Shurui and Ji, Shuiwang, Explainability in Graph Neural Networks: A Taxonomic Survey, arXiv, 2020,



Yuan, Hao and Yu, Haiyang and Gui, Shurui and Ji, Shuiwang, Explainability in Graph Neural Networks: A Taxonomic Survey, arXiv, 2020,

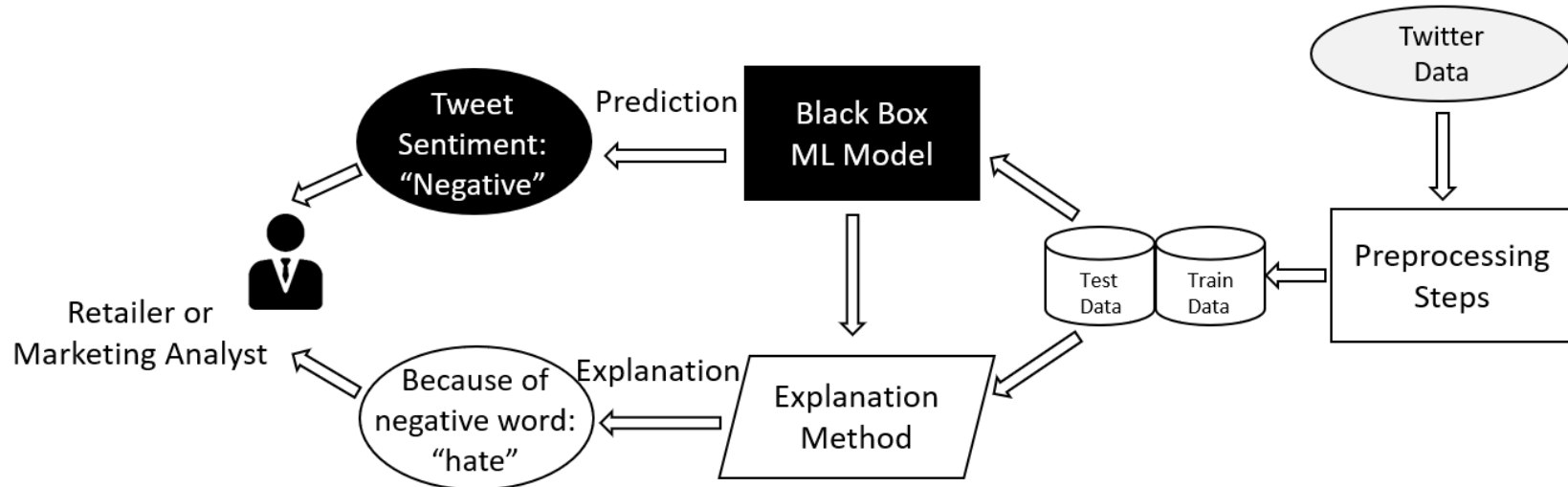
Partie 2.

Sentiment Analysis Application

Partie 2

Sentiment Analysis Application

- **Explainable sentiment analysis**
- **Our approach for Detecting opinion change**
 - **Form ego Networks to Propagation network**
 - **Dataset: topological and textual information**
 - **Machine Learning and explainability**



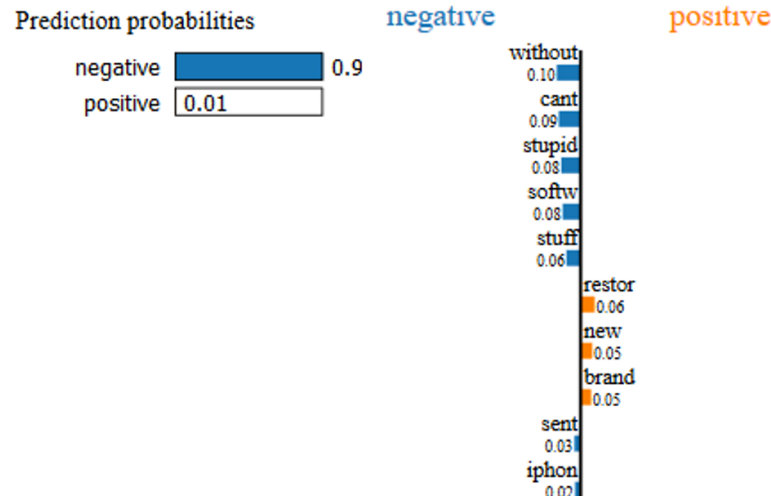
- Dataset of tweets discussing electronics products
- 375 positive and 1092 negative tweets,
- a total of 1467 instances.

| ML Model | Average F1 Score |
|----------|------------------|
| SVM | 0.75 |
| RF | 0.74 |
| XGBoost | 0.79 |
| MLP | 0.81 |

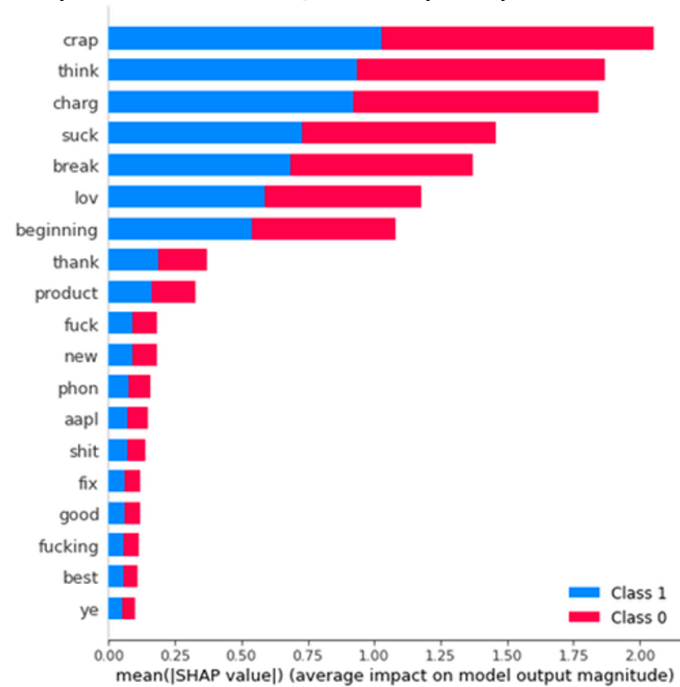
Cirqueira, Douglas et al. Explainable Sentiment Analysis Application for Social Media Crisis Management in Retail. *CHIRA* (2020).

Tweet: "strik somewhat stupid sent brand new iphon without latest softw cant restor old stuff"

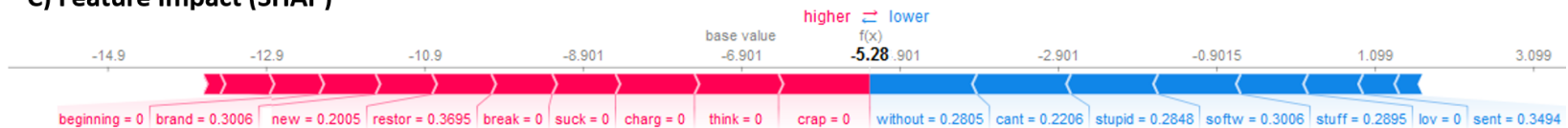
A) Local Feature Importance (LIME)



B) Global Feature Importance (SHAP)



C) Feature Impact (SHAP)



Partie 2.

Sentiment Analysis Application

Partie 2

Sentiment Analysis Application

- Explainable sentiment analysis
- **Our approach for Detecting opinion change**
 - Form ego Networks to Propagation network
 - Dataset: topological and textual information
 - Machine Learning and explainability

- Study the interaction between two ego networks around two influencers having opposite opinion on a given topic
 - its impact on opinion change and propagation within these two interconnected ego networks.
- **Goal:** explore the combination of sentiment analysis with complex networks analysis
- **Proposition: Explainable Method for detecting opinion modification** in relation to several nodal and topological measures: users' centralities, the opinion of the community to which belongs the users, as well as textual information extracted from tweets.

Folly, K., Malek, M., & Kotzinos, D. [Social networks analysis for opinion model extraction.](#)
In *Networks 2021: first combined meeting of the International Network for Social Network Analysis (Sunbelt XLI), and the Network Science Society (NetSci 2021)*, Indiana, United States, July 2021.

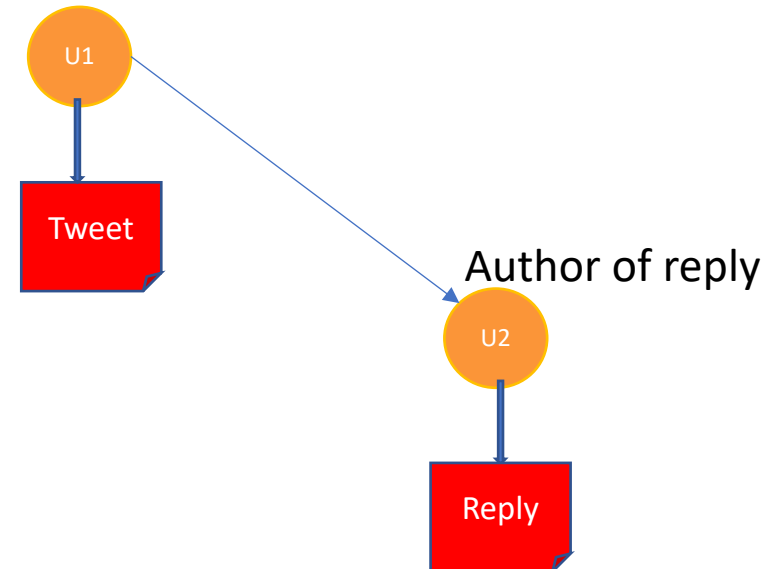
1. **Data collection** from twitter: retrieve tweets related the chosen topic
2. **Find two opposite influencers**
 - metrics: popularity, range the propagation, likeness scores, users' profiles
3. **Construction of the propagation directed network**
 - nodes: common repliers extracted from both influencers' egocentric networks at levels 1 and 2
 - edges: the actions of reply.
4. **Study of the network characteristics** :
 - centralities, community detection: Louvain algorithm
5. **Sentiment Analysis**
 - polarity and subjectivity computing for: users, ego networks, communities
6. **(Explainable) Machine Learning for detecting opinion change of users**

- **Machine Learning for detecting opinion change of users**
 - Dataset: an entry is related to a link of the propagation network
 - nodal and topological features for both the original tweet and the reply author nodes
 - textual features are extracted with the TF-IDF method
- 1. **Detecting opinion change over time**
 - ✓ the target variable indicates if the replier has changed his opinion polarity over Time
- 2. **Detecting opinion modification via the action of reply**
 - ✓ the target variable indicates if the opinion is modified via the action of reply
- Find the best ML method for both

| Features nature | Features |
|----------------------|---|
| Topological features | <p>Betweenness centrality, Out centrality, In centrality, Pagerank, (for original tweet user) community_original_tweet_user, community_author, community_mean_subjectivity_original_tweet_user, community_mean_polarity_original_tweet_user, in_same_community</p> |
| Nodal features | <p>mean_polarity_original_tweet_user, mean_subjectivity_original_tweet_user, subjectivity_original_tweet_text</p> |
| Tweet features | <p>tweet_stopwords_ratio, reply_stopwords_ratio (low, medium, high, very high), length_reply_high (low, medium, high, very high), length_tweet (low, medium, high, very high)</p> |

Without textual features

Original tweet

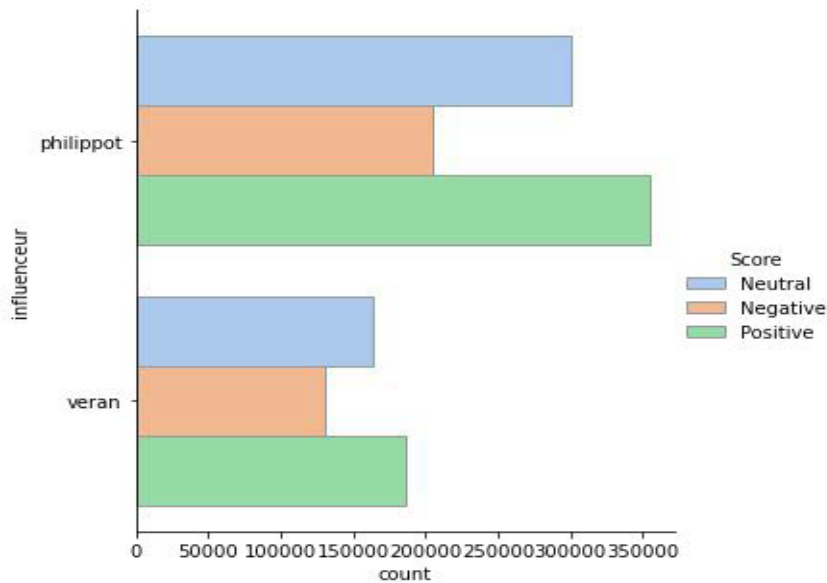


- The chosen topic **Covid vaccination**
- Period for retrieved tweets : October 1, 2021 to December 14, 2021.
- The query returned approximately 65.000 tweets.
- The found influencers were
 - Florian Philippot, politician against vaccination
 - Olivier Véran, for vaccination.

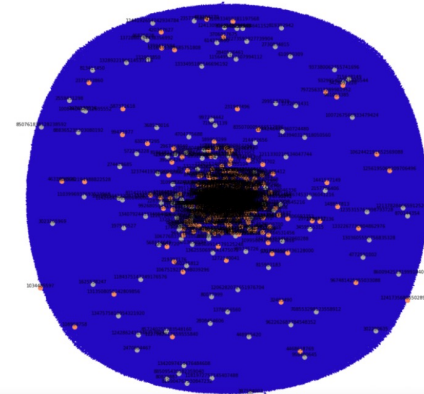
| Nodes # | edges # | Outdegree | Indegree | Betweenness | pagerank | Communities # |
|---------|---------|-------------|---------------|---------------|------------------|---------------|
| 21075 | 84441 | [0.05,0.18] | [0.001,0.003] | [0.002,0.007] | [0.00004,0.0005] | 36 |

Propagation network: characteristics (top 10)

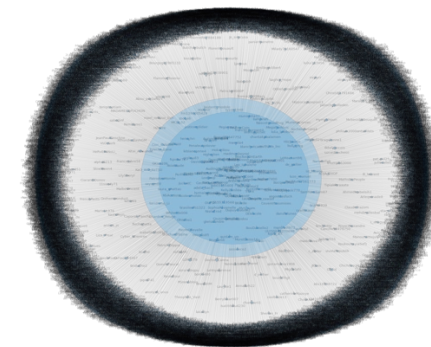
- **Out-degree centrality** have significant values
 - detect influencing users that contribute to the propagation of opinions



Opinion propagation in both egocentric networks

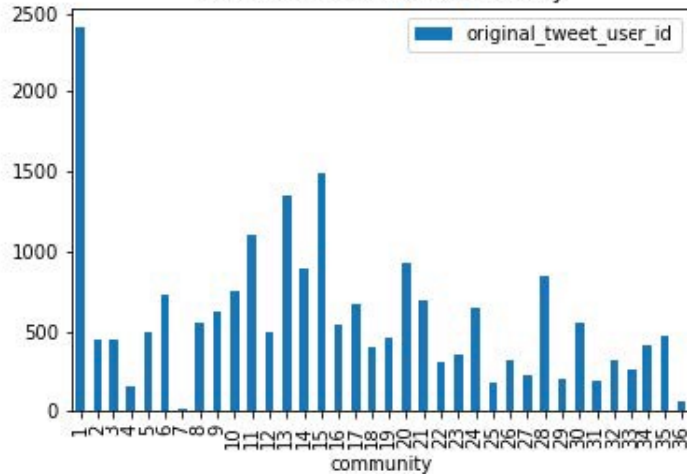


Propagation network

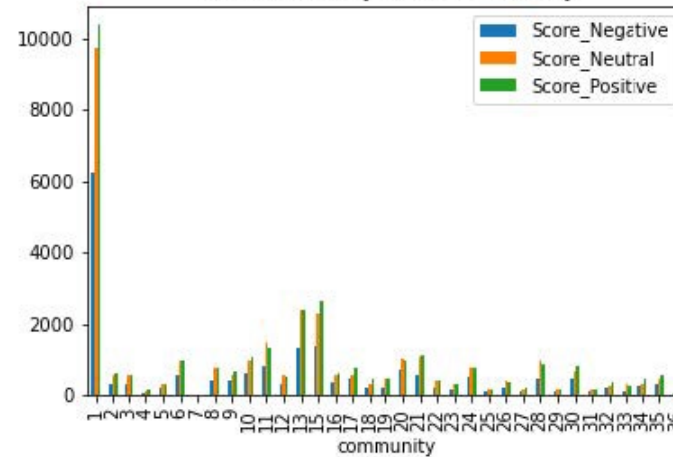


Retweet network

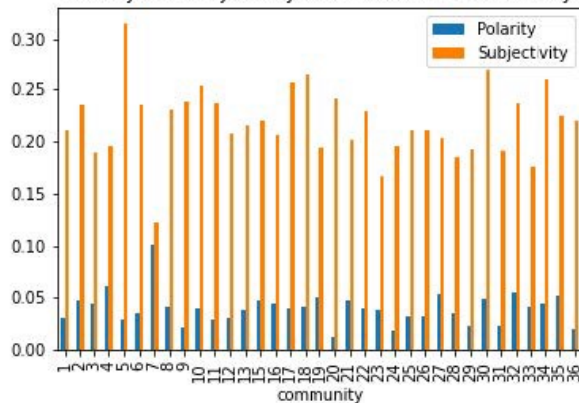
Number of Users Per Community



Sentiment Analysis Per Community



Polarity and Subjectivity Mean Scores Per Community

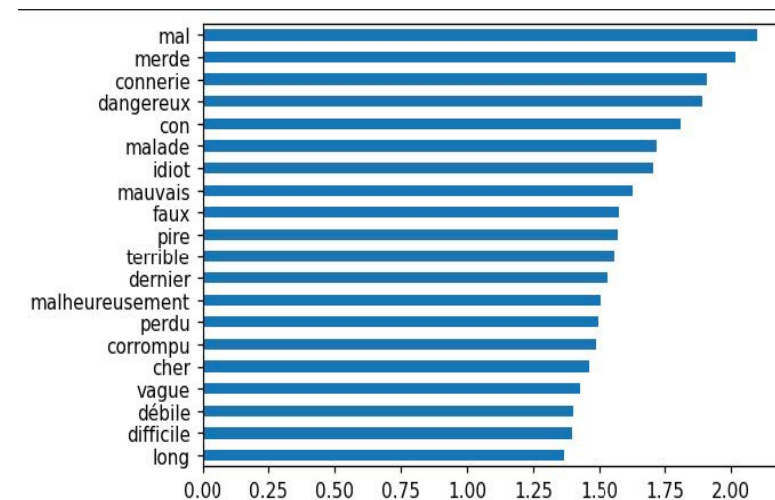
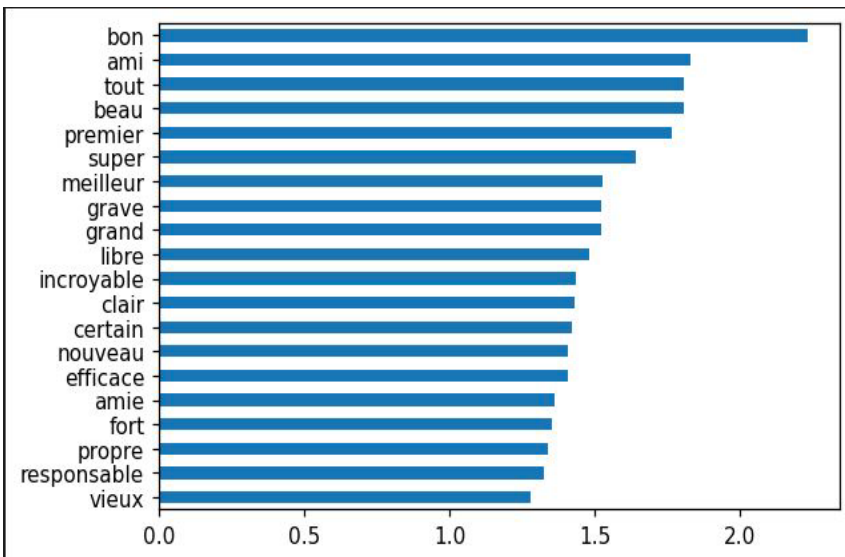


- ✓ **Polarities** are mixed and almost neutral within communities
- ✓ **No opinion polarization** within this configuration.

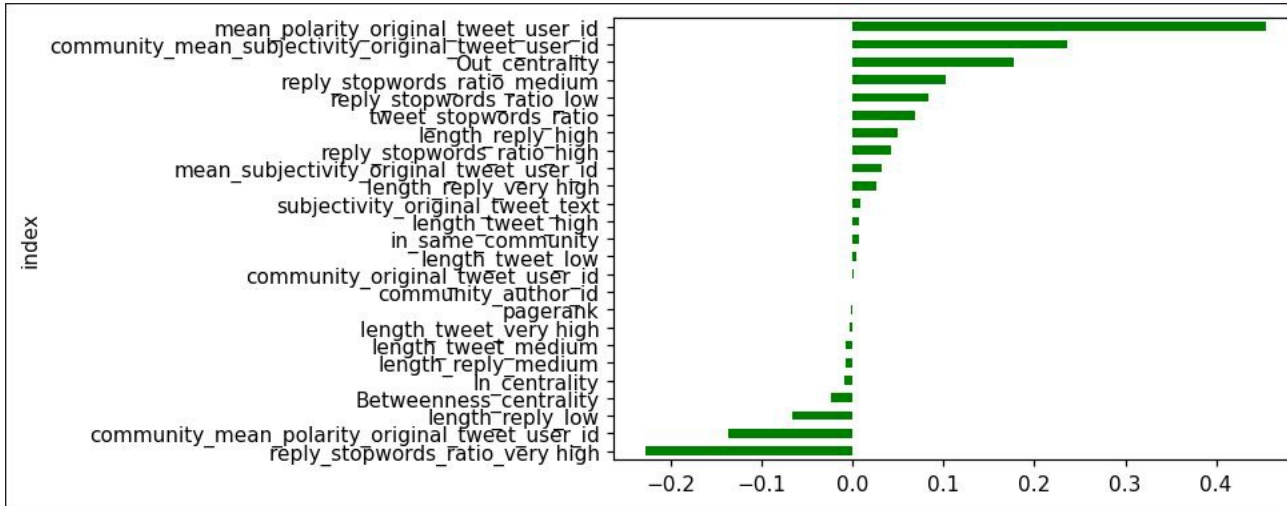
- Machine Learning for detecting user opinion change over time
 - Dataset: an entry is related to a link of the propagation network
 - nodal and topological features for both the original tweet and the reply author nodes
 - textual features are extracted with the TF-IDF method
- ✓ the target variable indicates if the replier has changed his opinion polarity over time

| Performance for the 3 class values | | | | | |
|------------------------------------|------------------|---------------|-----------------|-----------------|----------------|
| | <i>Precision</i> | <i>Recall</i> | <i>F1_score</i> | <i>Accuracy</i> | <i>Support</i> |
| TF-IDF | | | | 0.58 | 20062 |
| Increased | 0.67 | 0.64 | 0.65 | | 8249 |
| Decreased | 0.56 | 0.63 | 0.59 | | 8144 |
| Static | 0.44 | 0.36 | 0.38 | | 3669 |

Ridge Classifier

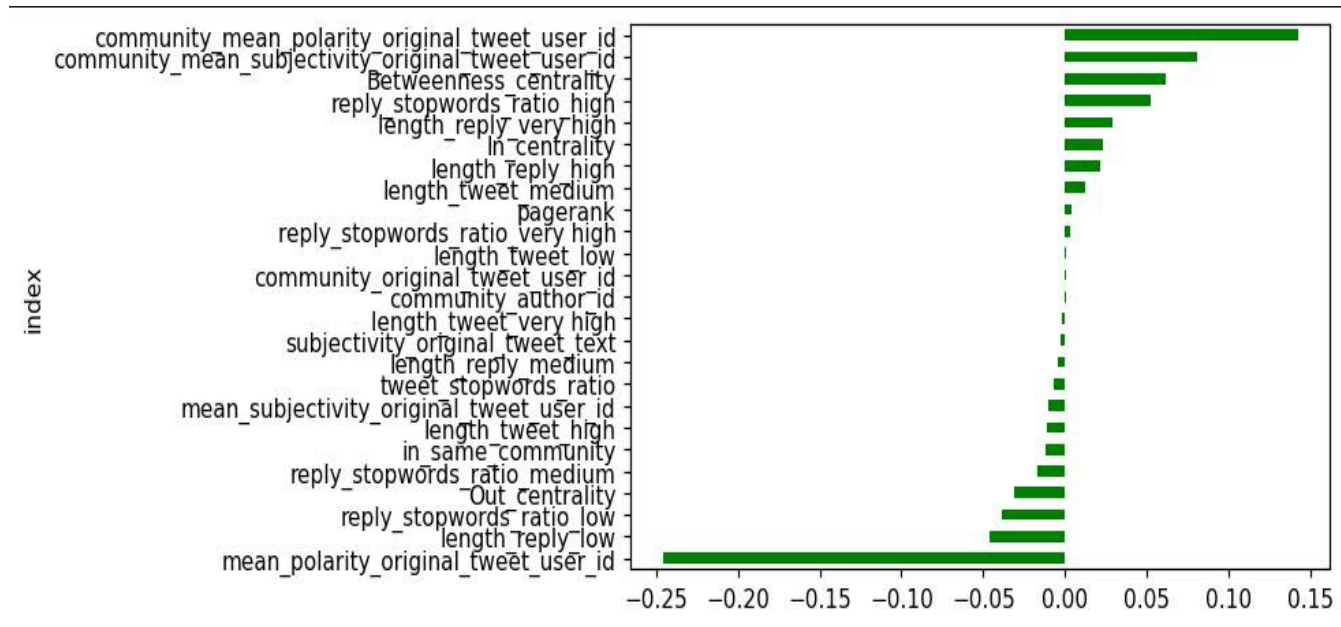


Ridge model coefficients for the increased (left) and decreased (right) class values : textuel coefficients.



Ridge model coefficients for the increased class value expect textual coefficients

the fact that the **origin user has a high polarity** in comparison to the polarity of its community, a **high out-degree centrality** and a **low betweenness one**, will contribute to the increase of the polarity of the replier.

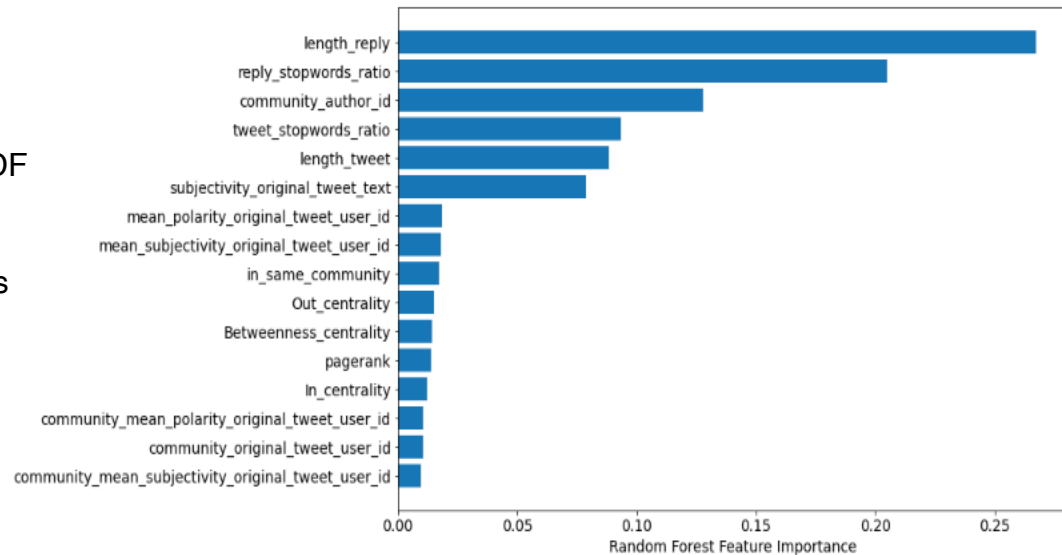


Ridge model coefficients for the decreased class value expect textual coefficients

➤ Machine Learning for detecting user opinion modification via the action of reply

- Dataset: an entry is related to a link of the propagation network
- nodal and topological features for both the original tweet and the reply author nodes
- textual features are extracted with the TF-IDF method
- ✓ the target variable indicates if the opinion is modified via the action of reply.

| Retained (best) Model : Random Forest | | | | | |
|---------------------------------------|------------------|---------------|-----------------|-----------------|----------------|
| | <i>Precision</i> | <i>Recall</i> | <i>F1_score</i> | <i>Accuracy</i> | <i>Support</i> |
| | | | | 0.59 | 20062 |
| True | 0.62 | 0.68 | 0.65 | | 11294 |
| False | 0.53 | 0.47 | 0.50 | | 8768 |



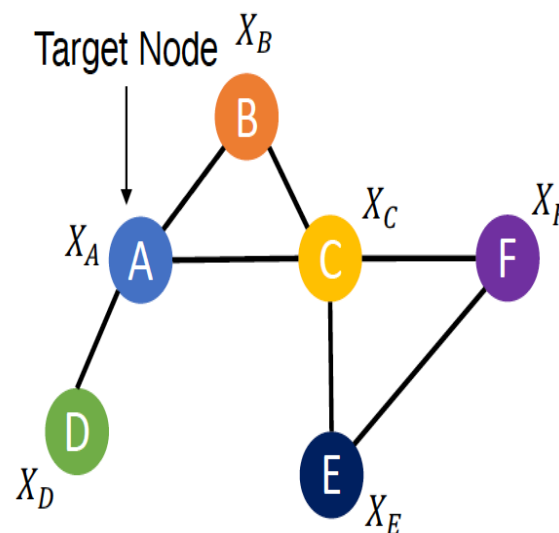
- Some variables concerning the form of the reply text have an important role in the detection of the opinion modification
- the community of the author (replier) seems to have also certain importance.

- Study the impact of the influencers ego networks' level (currently 2) on the polarisation of detected communities and the performances of opinion changedetection
- Study of the opinion formation in the propagation network models.
- Applying overlapping community detection algorithms
 - could help use to detect users that have role of moderators when spreading the information in the network.

Jean-Philippe Attal, Maria Malek, Marc Zolghadri. Overlapping community detection using core label propagation algorithm and belonging functions. *Applied Intelligence*, 51(11), 8067-8087, Springer,2021.

Explainability for interaction network algorithm ?

- Related Tasks: link prediction, node classification or searching, community detection, multilayer exploration, etc.
- Graph nature : similarity graph, interactions, social, etc.
- Can we distinguish **transparent models and opaque models**?
- Can we classify the **model Types**?
depth first research, breadth first research, random walk, shortest path, label propagation, heuristic optimization, etc.
- Can we define **explainability categories**? local, visual, by simplification, etc.
- **Explainability principles**? incidents links, neighbor nodes, attribute nodes, centrality measures, replay last steps of graph exploration, etc
- ..



Examples (Use cases)

➤ Opinion formation:

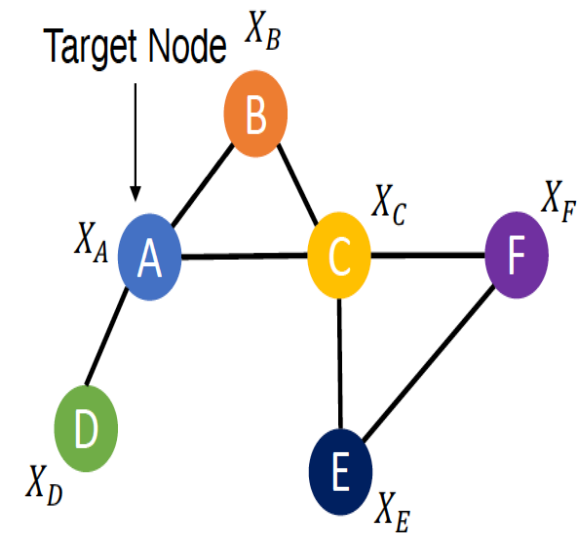
$$x_i(t + 1) = x_i(t) + \frac{\mu}{N_i} \sum [x_j(t) - x_i(t)]$$

➤ Recommendation:

- Recommend an item or user: exploring a part of the graph
- Recommend node because a given threshold is reached and explored edges have "height weights", heuristic is optimized, etc.

➤ How to explain?

- Node attributes and centrality measures
- Weight of incident links
- Information about neighbors (attributes + centrality measures)
- Replay last steps



[Dalia Sulieman](#), [Maria Malek](#), [Hubert Kadima](#), [Dominique Laurent](#):

Toward Social-Semantic Recommender Systems. [Int. J. Inf. Syst. Soc. Chang.](#) 7(1): 1-30 (2016)

Amir Mohammadinejad. Consensus opinion model in online social networks based on the impact of influential users. PhD thesis, Telecom & Management SudParis, Evry, France, 2018.



Thank you!

ETIS

Équipes Traitement
de l'Information
et Systèmes

Laboratoire Etis
6 Rue du Ponceau
95000 Cergy

etis-lab.fr

T.07 61 76 91 47

