Universidad del Desarrollo
Facultad de Gobierno
Centro de Investigación en Complejidad Social

TEXT-BASED LINK PREDICTION IN SOCIAL NETWORKS

POR: IGNACIO TOLEDO ROMÁN

Tesis presentada a la Facultad de Gobierno de la Universidad del Desarrollo para optar al grado académico de Doctor en Ciencias de la Complejidad Social

PROFESORES GUÍA:

Sr. JORGE FÁBREGA LACOA

Diciembre, 2022

SANTIAGO

DEDICATORIA

A mi esposa Belén, a mis hijos Sofía y León, y a mis padres Lubachka y Rodrigo

AGRADECIMIENTOS

Son muchas las personas sin las cuales esta tesis no habría sido posible y a ellas les expreso mi profunda gratitud. En primer lugar, a mi esposa Belén Arriagada cuyo amor y apoyo incondicional fue fundamental en este proceso. Ella siempre me ha motivado a perseguir mis objetivos y a no rendirme. Durante la realización de la tesis fue quién me acompañó y cuidó durante las noches en vela. A mis hijos Sofía y León, por su inmenso cariño y alegría, también por entender a su corta edad las ausencias y trasnoches. A mis padres, por siempre fomentar mi curiosidad, y desde joven, animarme a decidir sobre mis estudios. Su impulso fue vital para iniciar el doctorado. Son muchos los compañeros y compañeras del doctorado a quienes les guardo un gran cariño, en especial quiero agradecer a Mauricio Aspé, Carina Sturla, y María Teresa Barbato con quienes compartí interminables horas de estudio, por su amistad y compañía que fue crucial durante los primeros años del doctorado. Además, no puedo dejar de agradecer a mi amigo de la vida, Fabián Zambrano por estar siempre presente y por ayudarme a retomar la tesis en uno de los momentos más difíciles, cuando a causa de la pandemia me vi obligado a postergarla. También debo a agradecer a profesor tutor Jorge Fabrega, y los profesores de mi comisión Leo Ferres y Miguel Guevara, por sus comentarios y sugerencias, que de manera constructiva, me dieron valiosas herramientas, para seguir la carrera académica. Por último, quiero agradecer a los profesores y funcionarios del Centro de Investigación en Complejidad Social de la Universidad del Desarrollo que de alguna u otra forma han sido parte de este camino, y que de ellos aprendí el valor de la interdisciplina y el rigor científico.

TABLA DE CONTENIDOS

INTRODUCTION

1.Scientific collaboration networks

During the past decades, the scientific production process has transitioned from being a solitary activity to a more collaborative endeavor. This shift is reflected in an exponential growth of the proportion of papers published by teams relative single-authored papers (Abt, 2007). However, it isn't just the number of teams in science that has changed, the size of the teams in most fields is also increasing (Fortunato et al. 2018). The size of teams have also increased research impact, measured as the number of citations received (Wuchty et al., 2007). This is also true for research quality, papers authored by larger teams are more likely to receive a better quality rating by peer reviewers (Franceschet and Costantini, 2010).

Contemporary science can be represented as a large, complex, self-organizing, and evolving network of scholars, projects, papers, and ideas (Fortunato et al. 2018). Thanks to the increased access to large bibliographic datasets that capture major activities in science we are now able to analyze the traces of this dynamic interactions and use mathematical and computational models to explore patterns in scientific production (Zeng et al. 2017).

The increasing dominance of teams in scientific production and the explosive number of available data on collaborative interactions has drawn a great deal of attention to the study of scientific collaborations from the perspective of complex networks. In scientific collaboration networks, nodes represent scholars and two scholars are connected if they have co-authored a paper in the past (Newman 2001). Early studies in complex networks have contributed greatly to our current understanding of the structure and evolution of scientific collaboration networks, particularly describing properties like the small-world and scale free network structures, community structure and assortative mixing (Newman 2002, Barabási 2002, Newman 2003, Newman and Girvan 2004).

Being able to predict where and when the next scientific collaborations will occur generates unprecedented opportunities not only for scholars wanting to develop their careers but also for Universities, R&D centers, stakeholder in the innovation ecosystem and policy makers.

Given a state of a network, the task of identifying the pairs of nodes that are more likely to form new connections in a future state is called the link prediction problem (Liben-Nowell and Kleinberg 2007). This predictive approach allows us to assess theoretical models of link formation using real world data, and to detect missing link and unobserved relationships.

2. Text-based link prediction in scientific collaboration networks

The initial formalization of the link prediction problem for social networks raised asks: to what extent can the formation of new links be modeled relying solely on features derived from the network topology? (Liben-Nowell and Kleinberg 2007). As topology-based models do not depend on the context of the network, they can be freely applied to complex networks in many fields such as biological, technological and information networks (Lü and Zhou 2011, Martínez et al. 2017).

Many studies have shown that including non-topological features can significantly increase the performance of link prediction models (Hassan et al., 2006; Sachan and Ichise, 2010; Wang et al 2015). These findings have expanded the scope of the link prediction problem to include the underlying question: to what extent can the link prediction models be improved by the inclusion of domain-specific features?

However, the lack of systematization of the recent developments in text-based link prediction makes answering this question for the domain of scientific collaboration network. Since topology-based methods and applications are in constant evolution, recent surveys of link prediction in complex and social networks have paid little attention to text-based link prediction methods (Martínez et al., 2017; Pandey et al. 2019; Kumar et al., 2020; Mutlu et al. 2020; Samad et al. 2020; Yuliansyah et al., 2020).

Recent advances in text-based link prediction methods focus on sophisticated machine learning and deep learning methods, but using basic text-processing and attribute extraction techniques. In spite of the great availability of massive scientific databases rich in text metadata (Priem, Piwowar, and Orr, 2022), and the ease of access to first-rate natural language processing techniques (Roberts et al., 2013; Mikolov et al., 2013; Pennington et al., 2014), new research and developments in text-based attributes are rather scarce.

This thesis aims to contribute to the literature on text-based link prediction methods in scientific collaboration networks with two original works. First, in Chapter I we present a comprehensive survey on text-based link prediction methods in social networks. And second, in Chapter II we introduce a simple but effective approach to improve the performance of text-based link prediction models.

CHAPTER I: A Survey of Text-Based Link Prediction in Social Networks

1. Introduction

When we observe the topology of a given social network, are we looking at all its significant connections? If that were the case, how will it evolve? Most studies have addressed these questions studying the topology of social networks and their attributes. In human social networks, for theoretical and empirical reasons, texts have become an important new source of data for that purpose in recent years, and existing surveys have not been enough to cover new contributions. The purpose of this article is to fill in that gap.

A social network is an abstract representation of a system of social interactions. Social networks are composed of a set of nodes that represent social actors (e.g., individuals and organizations) and links that represent the interactions between these actors (Wang et al., 2015). This formal approach provides a powerful framework for discovering and understanding the structural and dynamic properties of social systems. In this paper, we use graphs and networks interchangeably, nodes and vertices interchangeably, and links and edges interchangeably.

In most empirical studies on social networks, interactions between actors can not be directly measured. Instead, they are elicited from ethnographies, surveys, sampled observations, proxy variables, or from digital traces on large-scale online social networks. This brings two major challenges for researchers: network incompleteness and network evolution.

First, network incompleteness may induce measurement errors. Social interactions are rarely completely observable, as some of them may go unnoticed, be intentionally hidden, or be simply unobservable. In consequence, statistics describing the network structure are susceptible to bias. Discovering missing links is an increasingly relevant task to deal with noisy and incomplete data that may affect statistics reliability.

Second, network evolution may be underpinned by hidden mechanisms. Social networks are snapshots of highly dynamic systems. New links and nodes constantly appear in the network and some existing ones are likely to disappear from the network. Being able to predict which links will form in the future can help understand the forces driving the evolution of a social network.

Both challenges can be addressed as link prediction problems. The link prediction problem was originally formulated by Liben-Nowell and Kleinberg (2004) raising the question: to what extent can the formation of new links be modeled relying solely on features derived from the network topology? Along with formalizing the link prediction problem, Liben-Nowell and Kleinberg (2007)

benchmarked different similarity-based measures for social networks against a random predictor baseline showing major increases in performance.

Shortly after, Hasan et al. (2006) proposed framing the link prediction problem as a binary classification task that can benefit from the existing supervised learning techniques. This approach allows using similarity-based metrics as features for training machine learning models. Leveraging this approach the authors introduced a radical innovation by combining non-topological node attributes with traditional similarity-based topological features showing a significant increase in the performance of link prediction methods in an attributed co-authorship network.

Since then, several studies have addressed the link prediction problem proposing new methods. Great efforts has been made on systematizing the literature on link prediction in social networks (Hasan and Zaki, 2011; Wang et al., 2015; Pandey et al. 2019; Samad et al. 2020; Yuliansyah et al., 2020) and in the broader field of complex networks (Lü and Zhou, 2011; Martinez et al., 2017; Kumar et al., 2020; Mutlu et al., 2020).

The earliest taxonomy of link prediction methods identified three categories: 1) methods based on node neighborhoods; 2) methods based on the ensemble of all paths; and 3) higher level approaches (Liben-Nowell and Kleinberg, 2007) . Lü and Zhou (2011) expanded the scope of link prediction classification including new developments in similarity-based methods, in methods based on maximum-likelihood and in probabilistic models.

In parallel Hasan and Zaki (2011) contributed including binary classification models, probabilistic models and linear algebra models. Later, Wang et al. (2015) proposed a generic link prediction framework for social networks distinguishing between similarity-based methods and learning-based methods. All three, Lü and Zhou (2011), Hasan and Zaki (2011) and Wang et al. (2015) recognize the benefits of including external information such as nodes attributes into the features set in learning-based methods, improving the link prediction performance.

In attributed networks, nodes are often associated with a rich set of non-topological features (Huang, Li, Hu and 2018). In most cases the node attribute values are available in the form of text (Wang et al., 2015; Cai et al., 2017; Samad et al., 2020). Therefore, the use of text-based features for link prediction is becoming an increasingly common practice, particularly for domain-specific practical applications.

In the last few years new developments on text-based link prediction methods for attributed social networks have taken place. However, recent surveys (Pandey et al. 2019; Samad et al. 2020; Kumar et al., 2020; Mutlu et al., 2020) concentrate their contributions mostly on updating topology-based methodologies. Although these surveys clearly state the benefits of using node attributes external to the network (e.g. demographic data, personal interests, skills and social behaviors) new methods in this matter are rarely covered.

Therefore, although there exists wide consensus about its benefits (Lü and Zhou, 2011; Hasan and Zaki, 2011; Wang et al., 2015; Martinez et al., 2017) the existing literature on link prediction methods based on non-topological attributes have two persisting weaknesses. First, a growing corpus of text-based link prediction methods has not been covered in recent surveys. Second, the issue of link prediction in attributed networks has not been comprehensively addressed in previous surveys. In this regard our research focuses on text-based link prediction. The contribution of this paper is twofold:

- In depth analysis of the link prediction problem for attributed networks.
- An updated review of new text based methods not covered in previous surveys

The paper is organized as follows. Section 2 describes the link prediction problem in attributed networks and introduces a generic framework for text-based link prediction. Section 3 addresses the different partition approaches and proposes a taxonomy for their classification. Alternative approaches for text data processing are reported in Section 4. Section 5 focuses on categorizing emerging text-based link prediction techniques for attributed networks. Prediction evaluation approaches are described in Section 6. Finally, conclusions and directions for future work are discussed in Section 7.

2.Link prediction problem for attributed networks

The simplest social network (i.e. homogeneous, unweighted, unsigned, undirected, and non-attributed) can be denoted by $G = \langle V, E \rangle$ where V represents the set of vertices, E is the set of edges. Similarly, attributed networks, can be denoted by $G = \langle V, E, X \rangle$ where X is the matrix of attributes of nodes.

The attributed network G can also be expressed as a $G = \langle A, X \rangle$ where $A \in \mathbb{R}^{n \times n}$ represents the corresponding adjacency matrix, and $X \in \mathbb{R}^{n \times m}$, where $n = |V|$ is the number of nodes in the network is n, and m is the number of attributes of the X matrix. Each row in X, denoted by $x_i \in \mathbb{R}^m$ is the array of attributes corresponding to the node $v_i \in V$ (see Figure 1).

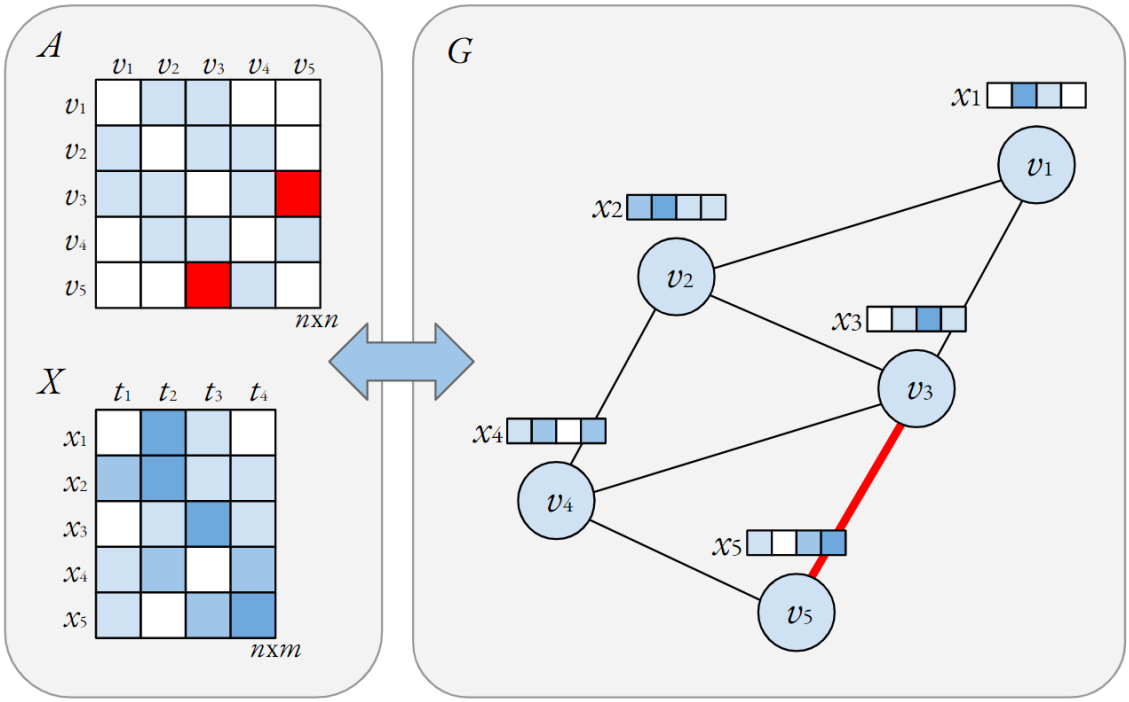


Figure 1: Representation of an attributed network with $n = 5$ and $m = 4$

The link prediction problem for attributed networks can be formalized as a binary classification task (Hasan et al., 2006). To do so, the network data is partitioned into two subgraphs. The first subgraph G_0 corresponds to the training set and contains the information that will be used to generate the prediction. The second subgraph G_1 corresponds to the testing set which provides the ground truth used for the final evaluation of the prediction. In practice, the number of the partitions, and the training-testing size ratio vary greatly depending on the partition approach and other methodological decisions addressed in Section 3.

Considering $u, v \in V$ two vertices of the attributed network G, the $y_1^{\langle u,v \rangle}$ array represents the label of each pair of vertices $\langle u, v \rangle$ in the testing partition G_1. In the case that the pair $\langle u, v \rangle$ exists in E_1 the label value is +1, if not it is labeled as -1. For simplicity it is assumed that the nodes set does not change $V_1 = V_0 = V$.

$$y_1^{\langle u,v \rangle} = \begin{cases} +1, \ if \ \langle u, v \rangle \in E_1 \\ -1, \ if \ \langle u, v \rangle \notin E_1 \end{cases}$$

Then, an algorithm P is granted with access to the training subgraph G_0 to perform the classification. The algorithm's output is a set of edges $E_p = P(A_0, X_0)$ that are predicted to appear in the network G_1 that are not contained in G_0.

Consequently, the list of predicted edges E_P should be a subset of the non-observed links $E_P \subseteq U - E_0$, where the universal set U contains all the possible $n \times (n-1)/2$ links, and E_0 is the set of edges of the subgraph G_0. The predicted edges E_P can also be represented as a label array $y_P^{\langle u,v \rangle}$.

$$y_P^{\langle u,v \rangle} = \begin{cases} +1, \ if \ \langle u, v \rangle \in E_P \\ -1, \ if \ \langle u, v \rangle \notin E_P \end{cases}$$

2.1. Text-based link prediction framework

In practice, link prediction techniques are embedded in a series of methodological steps that includes data wrangling, data partition, data processing and performance evaluation. Decisions made within each step, may affect not only the overall performance of link prediction tasks but also the interpretation of the role of node attributes in the final result.

Text-based link prediction methods may be particularly sensitive to this type of early stage data processing decisions since a wide variety of natural language processing techniques are constantly being developed.

Based on the work of Wang et al. (2015) we propose a generic framework for text-based link prediction methods that breaks down the link prediction workflow into four steps: data partition, text processing, link prediction, and performance evaluation (see Figure 2).

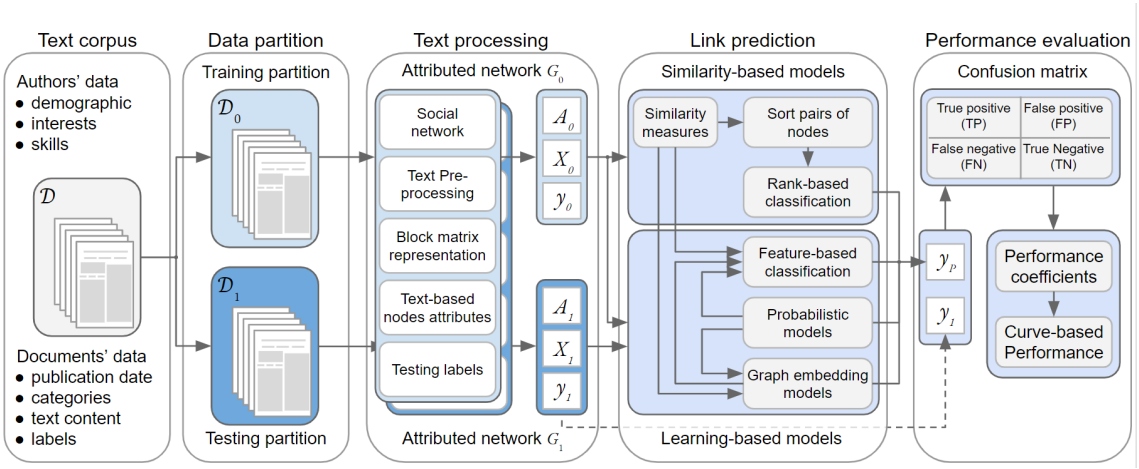


Figure 2: Generic text-based link prediction framework

The first step consists in choosing the appropriate partition approach that matches with the link prediction task, for example: deciding if the partition will be based on time interval or randomly selected. Different partition approaches are further described and categorized in Section 3.

The second step consists of mapping the text corpus $\mathcal{D}$ into the attribute matrix X. This means to process the text from the document collection to build an attribute profile for each node based on the text associated with it. In Section 4 we review different methods for calculating X.

The third step corresponds to the link prediction method. In the case of similarity-based methods a similarity measure is calculated from the training partition data giving each pair of nodes a similarity score $s(u, v)$ according to which top ranked pair of nodes are predicted to form a new link. The group of learning-based methods involves a wide range of supervised and unsupervised statistical methods that process the network structure and non-topological attributes to generate a prediction. In Section 5, text-based link prediction methods are described and classified.

The final step consists of assessing the performance of the generated prediction by comparing the predicted label array $y_P^{\langle u,v \rangle}$ to the testing label array $y_T^{\langle u,v \rangle}$ from the testing subgraph G_1. The comparison results are summarized in a confusion matrix which allows to estimate performance coefficients and curve-based performance indexes. Performance assessment approaches are further discussed in Section 5.

3.Partition approaches for link prediction

Link prediction typically addresses two problems: predicting new links that will form in a future state of a dynamic network (Liben-Nowell and Kleinberg, 2007), and discovering missing links in the current state of a static network (Clauset et al., 2008). The difference between both approaches is reflected on how data is partitioned. This categorization can be further divided according to the solution approach into unsupervised and supervised link prediction problems, as shown in Figure 3.

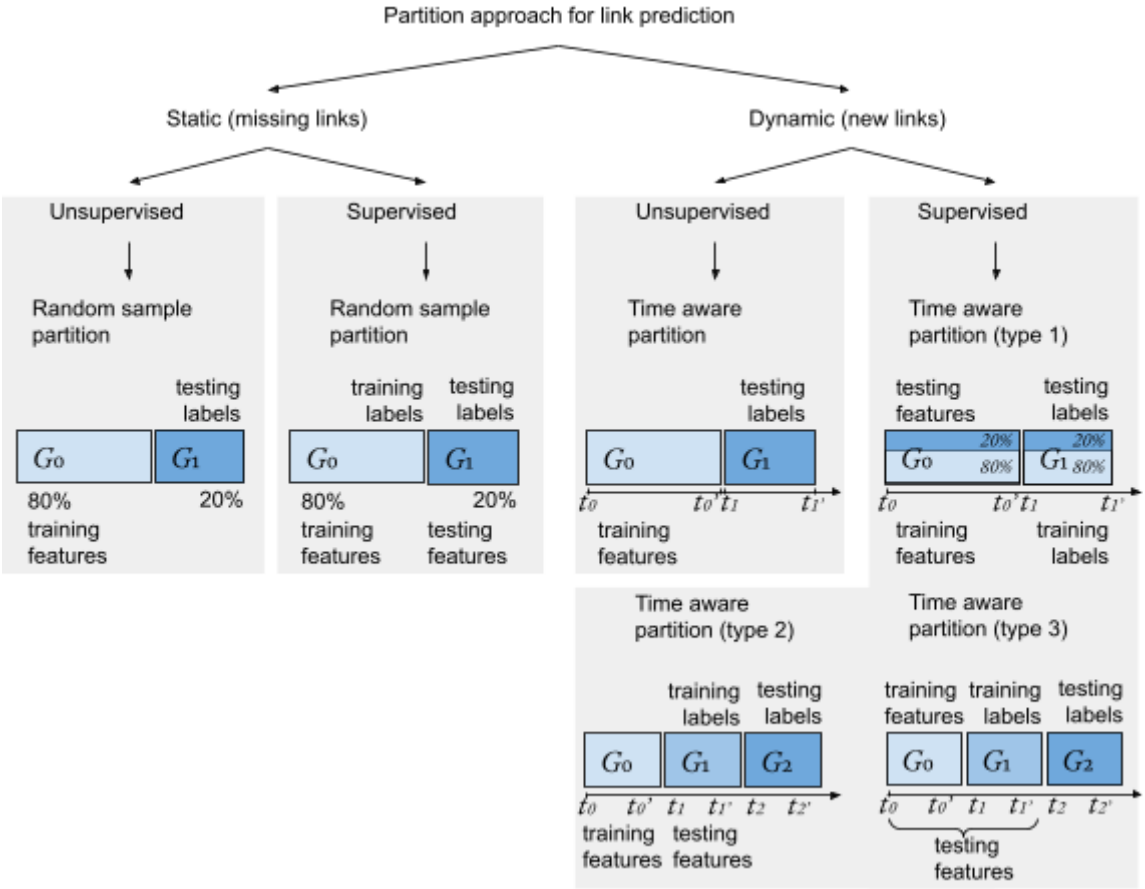


Figure 3: Taxonomy of partition approaches for link prediction

3.1.Missing links prediction in static attributed networks

Since the missing link approach does not depend on time, the partitions of data are defined by random sampling. The distribution of the size of partitions usually range from 90% to 80% of the dataset for the training subgraph G_0, and the remaining 10% to 20% for the testing subgraph G_1.

A common practice is cross validation of the link prediction results. For example: a 10-fold cross validation consists of dividing the dataset in 10 randomly sampled non-overlapping partitions, and for each one testing the link prediction algorithm using the remaining 90% of the data for training. The 5-fold cross validation is also widely used.

3.1.1. Unsupervised missing link prediction

Unsupervised methods for static networks are able to directly generate the set of predicted links $E_P = P(G_0)$ based solely on the information and features contained in the training subgraph G_0. This approach includes similarity-based metrics and unsupervised learning algorithms, for example k-means classification (Cai et al., 2017)

3.1.2. Supervised missing link prediction

In contrast to the unsupervised approach, algorithms based on supervised learning require labels as input, in order to estimate the model parameters. Therefore, training features derived from G_0 and the respective labels array $y_0^{\langle u,v \rangle}$ are used to fit the model. Once is properly trained the supervised model is provided with the testing features of the G_1 subgraph resulting in the predicted missing links $E_P = P(G_0, G_1)$.

3.2. Future links prediction in dynamic attributed networks

This approach considers time-aware data partitions based on link timestamps where each edge $e = \langle u, v \rangle \in E$ between a pair of vertices $u, v \in V$ occurs at a time $t(e)$, and for two points in time $t_n < t_n'$ the subgraph $G_n = \langle V_n, E_n, X_n \rangle$ consists of all edges that satisfy the condition $t_n \leq t(e) \leq t_n', e \in E$.

Then four points in time are chosen such that the condition $t_0 < t_0' < t_1 < t_1'$ is satisfied and the data is splitted into the subgraph G_0 corresponding to the training period $T_0 = [t_0, t_0']$ and the subgraph G_1 corresponding to the testing period $T_1 = [t_1, t_1']$.

3.2.1. Unsupervised future links prediction

Similarly to the missing links approach, the algorithm P directly generates the set of predicted links $E_P = P(G_0)$ based solely on the training features contained in G_0. The main difference with unsupervised missing link prediction is that G_0 is not chosen randomly but to satisfy the temporal partition condition for dynamic networks.

3.2.2. Supervised future links prediction

There are at least three types of time-aware partitions for link prediction methods based on supervised learning techniques.

- Type 1: Subgraphs G_0 and G_1 are further partitioned across both temporal partitions into two random samples with a defined ratio (e.g., 80% and 20%, 70% and %30). So, the model is trained using the information and features from the training subgraph $G_{0,80\%}$ and the labels $y_{1,80\%}^{\langle u,v\rangle}$ from the testing subgraph $G_{1,\,80\%}$. And the final prediction is obtained by feeding the fitted model with the information and features from $G_{0,20\%}$.

- Type 2: This approach considers a three partitions set-up. Subgraphs G_0 , G_1 and G_2, correspond to adjacent time periods $T_0 = [t_0, t_0']$, $T_1 = [t_1, t_1']$ and $T_2 = [t_2, t_2']$, such that $t_0 < t_0' < t_1 < t_1' < t_2 < t_2'$. Then, the features from G_0 and the labels $y_1^{\langle u,v\rangle}$ are used to train the algorithm, and the features from G_1 and the labels $y_2^{\langle u,v\rangle}$ to test its performance.

- Type 3: Similarly to type 2, this approach considers three periods T_0, T_1 and T_3. The training setup is exactly the same using the features from G_0 and the labels $y_1^{\langle u,v\rangle}$. But, for testing the prediction performance the features are elicited from the accumulated subgraph G_{0+1} corresponding to the period $T_0 + T_1$ and the labels $y_2^{\langle u,v\rangle}$.

Table 1: Partition approaches and training-testing ratios in text-based link prediction methods

Link prediction method		Partition approach			Training-testing partition ratio	
					Random*	Time-aware**
Makrehchi	2011	static	unsupervised		-	-
Liu et al.	2019	static	unsupervised		-	-
Zhang	2017	static	supervised		10-fold cross validation	-
Wang et al.	2017	static	supervised		10-fold cross validation	-
Hettige et al.	2019	static	supervised		80:20	-
Chaiwanarom et al.	2010	dynamic	unsupervised	type 1	-	8:2
Kong et al.	2016	dynamic	unsupervised	type 1	-	11:4
Solaimannhezad	2017	dynamic	unsupervised	type 1	-	6:7
Hasan, Chaoji, Alem and Zaki	2006	dynamic	supervised	type 1	5-fold cross validation	11:4

Wohlfarth and Ichise	2008	dynamic	supervised	type 1	90:10	3:3
Bartal, Sasson and Ravid	2009	dynamic	supervised	type 1	90:10, 80:20, 70:30	5:9
Sachan and Ichise	2010	dynamic	supervised	type 1	10-fold cross validation	6:2
Zhang and Yu	2014	dynamic	supervised	type 1	10-fold cross validation	2:2
Chuan et al.	2017	dynamic	supervised	type 1	10-fold cross validation	4:3
Hassan	2019	dynamic	supervised	type 1	10-fold cross validation	4:2
Ho, Bui and Bui	2019	dynamic	supervised	type 1	10-fold cross validation*	16:5
Rahmaida et al.	2019	dynamic	supervised	type 1	10-fold cross validation	5:5
Sun et al.	2011	dynamic	supervised	type 2	10-fold cross validation	10:7:7***
Wang et al.	2007	dynamic	supervised	type 3	-	8:1:1***

* Training-testing ratio in random partitions is expressed as percentages, for example 80:20 means 80% of the data is used for training and 20% for testing.

** Training-testing ratio in time-aware partitions is expressed as years, for example 8:2 means the first 8 years of the data is used for training and the last 2 years are used for testing.

*** In time-aware partitions types 2 and 3 the training-testing ratio is defined by three periods expressed in years.

4. Text data processing

In attributed networks, the node attribute matrix is derived from metadata available in the form of documents that may contain text or labels. The collection of documents is called corpus $\mathscr{D}$. Each individual document in the corpus can be related to one or more nodes.

An example of single-author documents are comments in online social networks such as Facebook, Twitter and LinkedIn. A good example of co-authored documents are scientific papers which are usually a product of collaborative research work.

The document content (i.e. text) consists of an ordered sequence of words and other symbols like spaces and punctuation. These words are interpreted in their relation with other words, and meaning is extracted from the text as a whole (Gentzkow, Kelly and Taddy, 2019).

In contrast, document labels consist of structured categorical information that may represent classification of individual documents in semantic domains, group or institutional affiliation, personal interests, among others (see Table 2).

Table 2: Types of metadata in attributed networks

Metadata	Data type	Node related examples	Document related examples	Representation approach
Texts	Unstructured	User biography, resume, or description	Comments in social media, titles and abstracts in academic papers	bag-of-words (frequency) word-embeddings
Labels	Structured	User interests, skills, or affiliations	Keywords, subject category, publication venues.	bag-of-words (occurrence)

Since the text in documents is inherently high-dimensional, a major challenge for text representation methods is to reduce the dimensionality of the data to manageable levels. Consequently, to deal with large volumes of data, text is often processed at term level. On the other hand, labels usually have narrower sets of terms making it easier to process the attributes.

The first step to estimate the attribute matrix is pre-processing the text to limit the size of the lexicon. Pre-processing techniques like removing stop words and stemming words can contribute to dimensionality reduction (Gentzkow, Kelly and Taddy, 2019).

Then, for text processing both authorship and text data are represented as a set of matrices. According to the text representation approach the processing techniques can be classified into two categories: bag-of-words representation techniques and word-embedding representation techniques. The resulting node-attributes of text processing are usually interpreted as a proxy of interest in certain topics for example, research interests, books, movies and sports.

4.1.Pre-processing techniques

A common practice to further reduce the lexicon size is filtering by "term frequency-inverse document frequency" $tfidf$. This measure represents the relevance of a word w_j in a document d_i, and is computed as the product $tf_{i,j} \times idf_j$ between the term frequency $tf_{i,j}$ and the inverse document frequency idf_j.

The term frequency $tf_{i,j}$ is equivalent to the count of the word w_j in a document d_i. In other cases, the count is normalized by the total word count, or the maximum frequency in the document. The inverse document frequency measures how common a word w_j is in the corpus and is computed as the log of one over the proportion of documents of the corpus containing the word w_j , $idf_j = log\ (|\mathcal{D}|/\mathcal{D}_j)$ where $\mathcal{D}_j = \left|d_i \in \mathcal{D}: d_{W_{i,j}} \neq 0\right|$.

This approach is widely used in the field of natural language processing for excluding too common and too rare words. Since it weights frequency and exclusivity, both rare words and common words will have lower $tfidf$ scores. A common practice is to keep only the words within each document i with $tfidf$ scores above some rank or cutoff (Gentzkow, Kelly and Taddy, 2019).

4.2.Block matrix corpus representation

The data contained in the corpus $\mathscr{D}$ can be represented as a set of matrices that describe relations between different entities like authors, documents, words, topics, among others. This set of individual matrices can be generalized as a block matrix (see Figure 4).

A similar approach is proposed by Huang et al. (2009) modeling some of these relations as a three-layer multi-network represented by an author-article-keyword block matrix. Makrehchi (2011) models author-topic relations as a semi-bipartite network represented as an author-topic block matrix. Sun et al. (2009) introduces the concept of heterogeneous information network that represents relations between authors, articles, keywords and publication venues as a block matrix (Sun et al., 2011).

For text based link prediction methods, the focus is on semantic relationships between authors, documents, words, labels, and topics. The relationships between authors, documents and words can be directly observed from the corpus or easily computed. On the other hand, latent topics are not directly observed and require to be inferred from the matrices that represent the corpus textual data.

We found the block matrix corpus representation useful to describe different approaches for text-based link prediction methods in a systematic way.

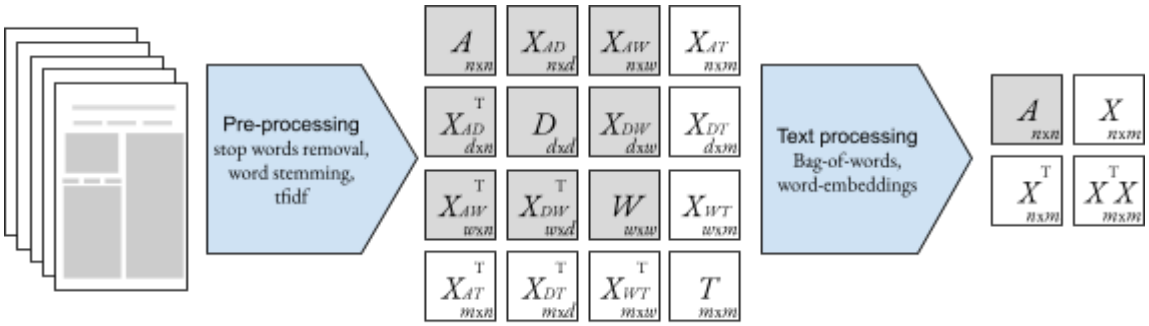


Figure 4: Block matrix representation of the corpus $\mathscr{D}$ for text processing.

As result of the text processing step, topic related submatrices are summarized into the node-attributed matrix X. The submatrices are described in Table 3.

Table 3: Definition of components of the block matrix representation of the corpus

Matrix	Entities	Description
A	Author-author	Given externally (e.g., friendship, followship), or computed as $X_D \cdot X_D^T$
X_{AD}	Author-document	Observed from the corpus authorship data
X_{AW}	Author-word	Computed as function of X_{AD} and X_{DW}
X_{AT}	Author-topic	Computed as function of X_{AD} and X_{DT}, or as function of X_{AW} and X_{WT}
D	Document-document	Given externally (e.g., citation, co-citation) or computed as $X_{AD}^T \cdot X_{AD}$
X_{DW}	Document-word	Observed from the corpus text data
X_{DT}	Document-topic	Computed from X_{DW} by using bag-of-words methods
W	Word-word	Given externally or observed from the corpus text data
X_{WT}	Word-topic	Computed from W by using word-embedding methods
T	Topic-topic	Given externally or computed as $X_{DT}^T \cdot X_{DT}$ or as $X_{WT}^T \cdot X_{WT}$

4.3.Bag-of-words representation

Bag-of-words representation is a widespread practice in fields like library science, information science and computer science. It is particularly popular in document classification tasks where it is commonly used to compute semantic features.

Modeling a document as a bag-of-words implies to ignore word sequentiality and their relative position obtaining a set of words without structure. This approach allows representing a collection of documents as a document-term matrix and performing algebraic operations and statistical analysis for dimensionality reduction. An advantage of this approach is its simplicity and low computational cost, while its downside is the loss of the rich syntactic information encoded in the sequence of terms. In link prediction, it has been used as semantic feature in combination with other topology-based features to train a feature-based classification model (Wang et al., 2007)

Each element $x_{DW_{i,j}}$ in the document-term matrix X_{DW} represents either the occurrence or the frequency of the word w_j in the document d_i. In the case of occurrence, $x_{DW_{i,j}}$ is binary value, and takes a one if w_j is present in the document d_i and receive a zero on the contrary. This approach is common when using texts with low word count (e.g., article titles) (Wohlfarth and Ichise, 2008; Rahmaida et al., 2019), or labels where terms can only occur once (e.g., keywords, subject categories, publication venue) (Hasan et al., 2006; Zhang, 2017).

In the case of frequency, $x_{DW_{i,j}}$ is a numeric value equal to the number of times w_j is present in the document d_i. Frequency approach is preferred in documents where the length of the document, the frequency of words and the size of the lexicon are less restricted (e.g. scientific papers, blogs) (Makrehchi, 2011, Hettige et al., 2019).

In both cases, the target is to estimate the document-topic matrix X_{DT}, that is a low dimensional representation of the documents of the corpus. And then, using the authorship information X_{AD}, calculate the author-topic attribute matrix X_{AT} (see Figure 5). Lowering the dimension of the attribute matrix contributes to more manageable data and less computationally intensive similarity measures.

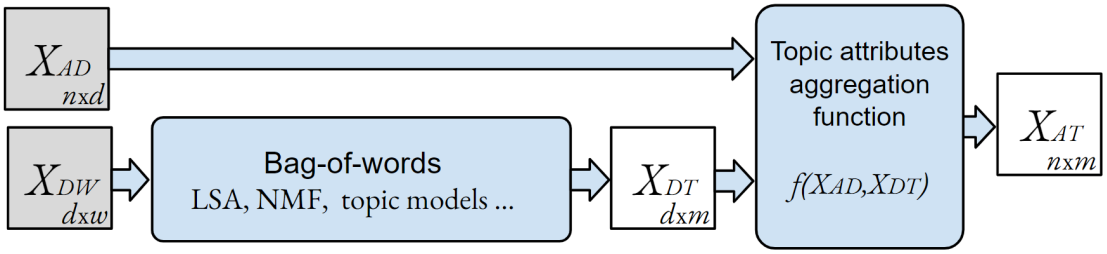


Figure 5: Diagram of a bag-of-words processing approach.

Common dimension reduction techniques for document-term matrices model the structure of the corpus discovering latent topics within the documents. These topics are constructs that group words with related meaning into intermediate levels between documents and words. This approach assumes that terms with related meanings will occur with similar distributions across the corpus' documents (Sahlgren, 2008). Dimension reduction techniques can be classified into two groups: algebraic methods and probabilistic methods.

Algebraic methods decompose the document-term matrix as the product of two or more matrices. Latent semantic analysis (LSA) is a widespread matrix factorization method and uses truncated singular value decomposition (SVD) keeping only the k largest singular values (Dumais, 2005). Non-negative matrix factorization (NMF) is another well known algebraic method that consists on decomposing the document-term matrix into to two matrices $M_{mxn} = W_{mxk} H_{kxn}$ with an intermediate dimension of k topics (Lee & Seung, 1999).

Probabilistic methods model the document-term matrix as a generative process where documents are described as probability distributions over k topics, those who in turn are defined as probability distribution over words. Hofmann (2000) introduced probabilistic latent semantic analysis (pLSA) which models each word in a document as a mixture of independent multinomial distributions that represent topics and learns the model parameters using the Expectation-Maximization algorithm. Latent Dirchlet allocation (LDA) is a generalization of pLSA where documents are modeled as a Dirichlet distribution; this feature allows estimating the parameters for new documents without overfitting problems (Blei, Ng and Jordan, 2003). A whole family of probabilistic topic models was built upon LDA, we list some of them:

- Author-topic model (ATM) (Rosen-Zvi et al., 2004): Estimates the representation of the content of each document in terms of topics, as well as the topical profile of each author .
- Correlated topic model (CTM) (Blei, Laffery, 2006): A central assumption in the LDA model is the independence of the topics. In contrast, CTM assumes a multinomial distribution instead of the Dirichlet distribution allowing correlations between topics.
- Supervised topic model (SLDA) (Chang and Blei, 2009): Jointly estimates the topic distributions and a response variable associated with each document, in order to find latent topics that will best predict the response variables for future unlabeled documents.

In section 5.2.1, we explore other probabilistic topic models applied to link prediction methods. Table 4 summarizes the text processing details in text-based link prediction methods using the bag-of-words representation.

Table 4: Text processing in bag-of-words approach

Link prediction method		Corpus	Metadata	Documents	Topic representation
Hasan et al.	2006	biobase and dblp	Labels	keywords	word occurrence
Wang et al.	2007	pubmed and dblp	Text	titles	tfidf
Wohlfarth and Ichise	2008	dblp	Text	titles	word occurrence
Bartal, et al.	2009	dblp	Labels	keywords	tfidf
Sachan and Ichise	2010	dblp	Text and Labels	titles, keywords and abstracts	word occurrence, tfidf
Makrehchi	2011	private corpus	Text	abstracts	word frequency, LSA, and LDA
Sun et al.	2011	dblp	Text	titles	word occurrence
Zhang and Yu	2014	medline	Text and Labels	keywords and abstracts	tfidf
Chaiwanarom and Lursinsap	2015	scopus	Text	titles, keywords and abstracts	ATM
Chuan et al.	2017	hep-th, hep-lat and amc	Text	titles and abstracts	LDA
Solaimannhezad	2017	private corpus	Text	abstracts	LDA
Zhang	2017	wos	Labels	keywords, categories, journals	word occurrence
Wang et al.	2017	arxiv	Text	abstracts	word frequency
Hassan	2019	arnetminer	Labels	keywords	word frequency
Ho, Bui and Bui	2019	dblp	Labels	keywords	LDA
Liu et al.	2019	private corpus	Labels	keywords	word occurrence
Rahmaida et al.	2019	scopus	Text	titles	word occurrence
Hettige et al.	2019	cora-ml, cora, citeseer, dblp, pubmed, and acm	Text	abstracts	tfidf

4.4.Word-embedding representations

In contrast with bag-of-words, word-embedding representation of text partially preserves word context and sequentiality leveraging the rich syntactic information. Word-embedding methods are used to process the word-word co-occurrence matrix W and learn models where word meaning is represented as a point in a m-dimensional vector space.

Meaning similarity between two words result in a close proximity in the vector space. This type of representation allows algebraic operations on word vectors that conserve syntactic and semantic regularities, for example the resulting vector of the operation *king-man+woman* should lie in close proximity to the vector *queen*.

The output of word embedding techniques is a set of word vectors represented as the word-topic matrix X_{WT}. Similarly to bag-of-words techniques, the dimensions in the vector space can be interpreted as representation of latent topics. The results are then aggregated for each author obtaining the author-topic attribute matrix X_{AT} (see Figure 6).

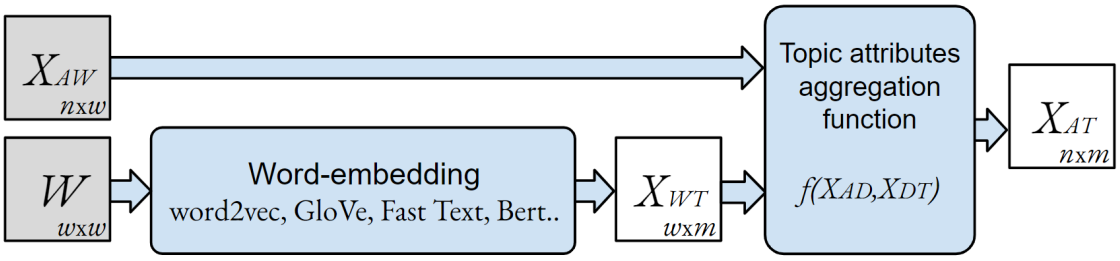


Figure 6: Diagram of word-embeddings processing approach.

Since the vector representation of each word contains both meaning and context information, word embedding models are particularly suitable for modeling short documents, such as comments in online social networks, (e.g. Facebook and Twitter) (Bhattacharyya et al., 2011; Amiri and Shobi, 2017; Baek and Chung, 2020) and scientific articles' titles and abstracts (Wu et al., 2018; Brochier et al 2019).

During the last decade several efficient word-embedding algorithms have been implemented using deep learning and other statistical techniques. These algorithms are able to solve text classification tasks with high accuracy and train models in reasonable time with accessible hardware specifications (D'Sa, et al. 2020; Dharma, et al. 2022). The open source implementations in programming languages like Python and R, have contributed to their growing popularity. We list some of them.

- Word2vec: Skip-gram (Mikolov et al., 2013)
- Word2vec: Continuous bag-of-words (CBOW) (Mikolov et al., 2013)

- Global Vector for Word Representation (GloVe) (Pennington et al., 2014).
- Fast text (Joulin et al., 2016)
- StartSpace (Wu et al., 2018)
- Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018)

Table 5 summarizes the details of text processing of text-based link prediction methods using the word-embeddings representation.

Table 5: Text processing in word-embeddings approach

Link prediction study		Corpus	Metadata	Documents	Topic representation
Bhattacharyya et al.	2011	facebook	Label	keywords	wordnet
Kong et al.	2016	dblp	Text	titles	word2vec
Amiri and Shobi	2017	twitter	Text	tweets	doc2vec
Wu et al.	2018	dbpedia, yelp, freebase	Text	title, abstracts, reviews, triplets	word2vec, GloVe, fasttext, StarSspace
Brochier et al	2019	cora	Text	abstracts	GloVe
Baek and Chung	2020	twitter	Text	tweets	word2vec, sentiment dictionary
Yoon et al.	2021	USPTO patent database	Text	patents	doc2vec

5.Text-based link prediction techniques

Link prediction techniques have rapidly evolved into a wide variety of methods. The classification of these techniques have also become an issue by itself. Previous surveys on link prediction in complex networks have proposed closely aligned taxonomies (Lü and Zhou, 2011; Martinez et al., 2017, Kumar et al., 2020; Mutlu et al., 2020, Yuliansyah et al., 2020).

In order to classify existing text-based link prediction methods we propose a two-level taxonomy compatible with previous efforts which is depicted in Figure 7. According to this approach, in the first level link prediction methods are classified into two main groups: Similarity-based and learning-based techniques. In the second level, both groups are further divided into subgroups.

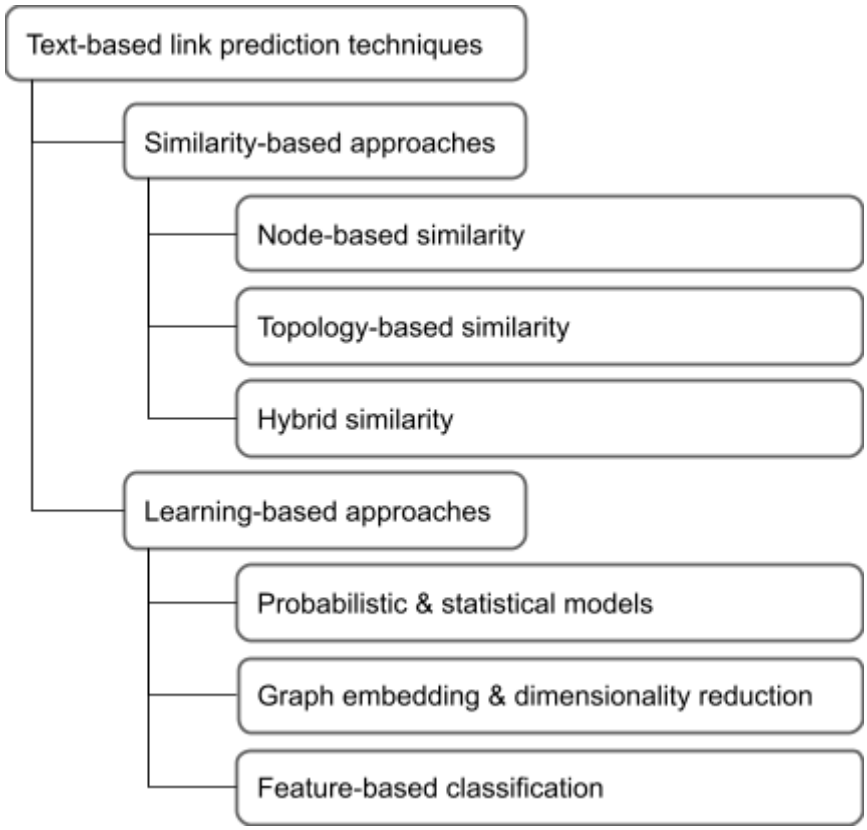


Figure 7: Taxonomy of link text-based prediction techniques

In addition, some link prediction approaches developed different interpretations of the attribute matrix X and their relations to the nodes. For each subgroup the proposed taxonomy identifies how node attributes are integrated into link prediction algorithms. The details of each reviewed method are described in the following sections.

5.1.Similarity-based models

The similarity-based measures are the foundation stone of most of current link prediction models. Similarity between nodes is modeled as closeness according to a given distance function or proximity measure. And similarity-based link prediction methods built upon the assumption that two nodes are more likely to form a new link in the future if they are similar to each other. Such versatile definitions have generated a wide variety of techniques traditionally rooted on topological attributes derived from the network structure (e.g. Common neighbors, Jaccard coefficient, Adamic-Adar, Katz index, etc.) (Liben-Nowell and Kleinberg, 2007).

Text-based similarity measures exploit this versatility in three different ways. One of them is by directly calculating the distance between nodes based on their text-based attributes (Hasan et al., 2006; Bartal, Sasson, and Ravid, 2009; Liu et al., 2019). In addition, topology based measures can be easily adapted to different types of networks like bipartite, semi-bipartite and heterogenous networks allowing the incorporation of node attributes as different types of nodes (Hasan et al., 2006; Makrehchi, 2011; Sun et al., 2011). Both approaches can be further combined into hybrid similarity measures (Chuan et al., 2017). Similarity-based measures are divided into these three groups, namely: node-based, topology-based, and hybrid similarity measures.

Similarity-based methods calculate a similarity score for each non-connected pair of nodes and then produce a ranked list in decreasing order. Top ranked pairs of nodes are supposed to be more likely to form new links.

5.1.1.Node-based similarity measures

The node-based techniques use closeness in a vector space as a proxy of node similarity. In this approach, the node-topic attribute matrix X_{AT} is interpreted as a collection of individual node attribute vectors $\{x_u = (x_{u,1}, x_{u,2}..., x_{u,m}) : u \in V\}$ that represent points in a m-dimensional vector space. Therefore, the similarity between a pair of nodes (u, v) can be modeled as the proximity between the vectors $x_u = (x_{u,1}, x_{u,2}..., x_{u,m})$ and $x_v = (x_{v,1}, x_{v,2}..., x_{v,m})$ in $\mathbb{R}^m$. This approach is widespread and traditionally used in the field of information retrieval (Salton et al., 1975).

The output of node-based measures is a similarity score that can be used as a sorting variable in rank-based predictions, and as a training feature in classification-based predictions. Similarity scores can also be used in conjunction with features obtained from learning-based techniques like graph embedding, and dimensionality reduction, like matrix factorization, deep learning, and graph kernel (Cai et al., 2017).

Manhattan Distance (D1): L_1distance, also called Manhattan distance, measures the sum of the lengths of each dimension of the difference of two vectors. It is commonly used in data analysis

(Phillips, 2021). Sachan and Ichise (2010) calculated X as the $tfidf$ matrix and used the L_1 distance to generate a semantic feature for training a binary classifier.

$$S_{D1}(u,v) = \left|\left|x_u - x_v\right|\right|_1 = \sum_{i=1}^{m} \left|x_{u,i} - x_{v,i}\right|$$

Euclidean Distance (D2): L_2distance, better known as Euclidean distance, is the most common distance measure. It is easily interpreted as the "straight-line" distance between two points or vectors (Phillips, 2021). Yoon and Magee (2018) predicted links between patents using this measure representing the patent's text-based attributed X_{AT} as the $tfidf$ matrix.

$$S_{D2}(u,v) = \left|\left|x_u - x_v\right|\right|_2 = \sqrt{\sum_{i=1}^{m} (x_{u,i} - x_{v,i})^2}$$

Cosine Similarity (COS): The cosine similarity measures the cosine of the angle between vectors x_u and x_v. This measure has been applied to a wide range of text representations for estimating the node-topic attribute matrix X_{AT}. Both, Hasan (2006) and Liu et al. (2019) built X_{AT} directly from the bag-of-words. In Bartal, Sasson, and Ravid (2009), Zhang and Yu (2014), and Wang et al. (2017) the topical attributes were represented as the $tfidf$ matrix. Chaiwanarom et al. (2015) estimated X_{AT} as the probability distribution of topics obtained from an ATM model. Duricic et al. (2021) used NMF approach to obtain a low dimensional attribute representation music interest of users in an online music platform (i.e., Last.fm).

The cosine similarity has also been used with word-embedding representations. Baek and Chung (2019) estimated X_{AT} using word2vec. Similarly, Yoon et al. (2021) used the cosine similarity between patents by estimating X_{AT} based on the doc2vec embedding method.

$$S_{COS}(u,v) = \frac{x_u \cdot x_v}{\left|\left|x_u\right|\right| \cdot \left|\left|x_v\right|\right|} = \frac{\sum_{i=1}^{m} x_{u,i} \cdot x_{v,i}}{\sqrt{\sum_{i=1}^{m} x_{u,i}^2 \cdot \sum_{i=1}^{m} x_{v,i}^2}}$$

In most studies the cosine similarity improved the overall classification performance and ranked among the most critical features in the link prediction task (Wang et al., 2007; Bartal, Sasson, and Ravid, 2009; and Zhang and Yu, 2014). Aiello et al. (2012) compared different similarity measures for link prediction tasks in online social networks data (i.e. aNobii, and Last.fm) and observed that cosine similarity performed the best. Zhang and Yu, (2014) observed that cosine similarity between $tfidf$ attributes was the second best rated among 12 topological and non-topological features.

Pearson Correlation Coefficient (PC): The Pearson coefficient measures the linear correlation similarity between two sets of data. Although the interpretation of the coefficient differs from the cosine similarity the difference between both measures is geometrically equivalent to a translation of the origin to the arithmetic mean values of the vectors (Egghe and Leydersdoff, 2009).

$$S_{PC}(u,v) = \frac{cov(x_u, x_v)}{\sigma(x_u)\cdot\sigma(x_v)} = \frac{\sum_{i=1}^{m}\left(x_{u,i} - x_u^{avg}\right)\cdot\left(x_{v,i} - x_v^{avg}\right)}{\sqrt{\sum_{i=1}^{m}\left(x_{u,i} - x_u^{avg}\right)^2 \cdot \sum_{i=1}^{m}\left(x_{v,i} - x_v^{avg}\right)^2}}$$

where x_u^{avg} denotes the average of x_u. A higher coefficient suggests a larger interest overlap between the two users. Huang et al. (2020) computed this coefficient to measure the similarity between the topic distributions of two users obtained from a LDA model estimated from user generated content in an online social network. The prediction performance of the coefficient was not reported independently, instead the results were used in combination with other indices to train different learning-based classification methods and were jointly assessed.

Hellinger Distance: This measure is used to quantify the similarity between two probability distributions (Hellinger, 1905). For discrete distributions the Hellinger distance is equivalent to the Euclidean distance between the squared roots of both distributions. Ho, Bui, and Bui (2019) used this measure to predict co-authorships based on the topic distributions obtained from probabilistic topic models (i.e., LDA and ATM).

$$S_{HD}(u,v) = \frac{1}{\sqrt{2}}\left\|\sqrt{x_u} - \sqrt{x_v}\right\|_2 = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{m}\left(\sqrt{x_{u,i}} - \sqrt{x_{v,i}}\right)^2}$$

5.1.2. Topology-based similarity measures

Topology-based techniques represent similarity between two nodes as a measure of how closely connected they are in the network. This approach considers text-based node attributes as entities part of the network. Consequently, the node-topic attribute matrix X_{AT} is interpreted as a component of an adjacency matrix.

In this section, we extend the traditional classification of topology-based similarity measures by including bipartite, semi-bipartite and heterogeneous networks (see Figure 8). Link prediction surveys have reported topology-based measures for bipartite and heterogeneous networks in the past, but as a case of variations of the link prediction problem (Kumar et al., 2020). In contrast, we analyze those cases from the perspective of the role of text-based attributes.

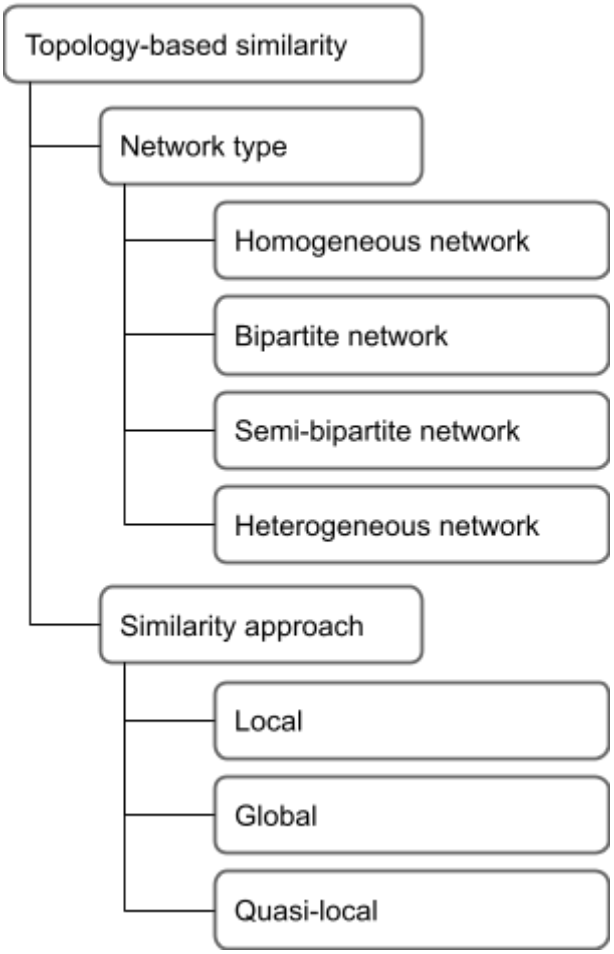


Figure 8: Classification of topology-based similarity measures by type of network and approach

5.1.2.1. Homogeneous networks

Topology-based similarity measures for homogeneous networks have been thoroughly covered by previous surveys (Wang et al., 2015; Martinez et al., 2017; Pandey et al., 2019; Kumar et al., 2020; Samad et al., 2020; Mutlu et al., 2020; Yuliansyah et al., 2020). These methods are typically classified in three main categories: Local, Global and Quasi-global measures.

We won't go into details about each individual method, for more information we recommend consulting Martínez et al. (2017). In Tables 6, 7 and 8, we present a rough summary of available techniques for each category. The equation, computational complexity, and reference of each algorithm are retrieved from previous surveys (Martinez et al. 2017; Kumar et al., 2020). A more detailed list of topology-based similarity methods can be found in the Appendix 1.

Local approach

Local similarity-base measure between a pair of nodes $\langle u, v \rangle$ are built based on the nodes' neighborhoods, $\Gamma(u)$ and $\Gamma(v)$ respectively. This approach has the advantage of being computed efficiently and highly parallelized (Martinez et al., 2017). Since the measurements are highly dependent on the intersection and union of both neighborhoods, its main drawback is that it can only be calculated for pairs of nodes at a distance of two steps. In Table 6 we present some of the most common methods for local similarity-based measures.

Table 6: Local similarity-based measures for homogeneous networks

Similarity measure	Equation	Time complexity	Reference
Common Neighbors (CN)	$S_{CN}(u,v) = \lvert\Gamma(u) \cap \Gamma(v)\rvert$	$O(vk^3)$	Liben-Nowell and Kleinberg, 2007
Jaccard Coefficient (JC)	$S_{JC}(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\cup\Gamma(v)\rvert}$	$O(vk^3)$	Jaccard, 1901
Adamic Adar (AA)	$S_{AA}(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{log\lvert\Gamma(z)\rvert}$	$O(vk^3)$	Adamic and Adar, 2004
Resource Allocation (RA)	$S_{RA}(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{\lvert\Gamma(z)\rvert}$	$O(vk^3)$	Zhou et al., 2009
Preferential Attachment (PA)	$S_{PA}(u,v) = \lvert\Gamma(u)\rvert \cdot \lvert\Gamma(v)\rvert$	$O(vk^2)$	Barabási and Albert, 1999
Sørensen Index (SO)	$S_{SO}(u,v) = \frac{2\cdot\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\rvert+\lvert\Gamma(v)\rvert}$	$O(vk^3)$	Sørensen, 1948
Salton Index (SI)	$S_{SA}(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\sqrt{\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert}}$	$O(vk^3)$	Salton and McGill, 1983
Leicht-Holme-Newman Index (LHN)	$S_{LHM}(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert}$	$O(vk^3)$	Leicht et al., 2006

Global approach

Global similarity-based measures are calculated using the whole network structure by algebraic operations on the adjacency matrix A (e.g. eigenvalues, pseudoinverse laplacian matrix). These approaches measure simultaneously the similarity between each possible pair of nodes without the restriction of the two-step distance. However, it comes with a higher computational complexity restricting scaling to large networks (Martinez et al., 2017). In Table 7 we present some of the most common methods for global similarity-based measures

Table 7: Global similarity-based measures for homogeneous networks

Similarity measure	Equation	Time complexity	Reference
Katz	$S_{Katz}(u,v) = \sum_{l=1}^{\infty} \alpha^l \lvert path_l(u,v) \rvert$	$O(v^3)$	Katz, 1953
Global Leicht Holme Newman Index (GLHN):	$S_{GLHN}(u,v) = D^{-1}\left(I - \frac{\alpha A}{\lambda_1}\right)^{-1} D^{-1}$ where λ_1 is the largest eigenvalue of adjacency matrix A and α is a free parameter.	$O(cv^2k)$	Leicht et al., 2006
Random Walk with Restart (RWR):	$S_{RWR}(u,v) = q_{uv} + q_{vu}$ where $\vec{q_u} = (1-\alpha)(1-\alpha P^T)^{-1}\vec{e_u}$	$O(cv^2k)$	Tong et al., 2006
Pseudoinverse Laplacian Matrix (PLM):	$S_{PLM}(u,v) = L^+_{uv}$ where L^+ is the Pseudoinverse Laplacian Matrix calculated as $L^+ = V\Sigma^+U^T = \left(L - \frac{ee^T}{n}\right)^{-1} + \frac{ee^T}{n}$	$O(v^3)$	Fouss et al., 2007
Pseudoinverse Laplacian Matrix Cosine (PLC):	$S_{PLC}(u,v) = \frac{L^+_{uv}}{\sqrt{L^+_{uu}L^+_{vv}}}$ where L^+ is the Pseudoinverse Laplacian Matrix	$O(v^3)$	Fouss et al., 2007
Average Commute Time (ACT):	$S_{ACT}(u,v) = \frac{1}{L^+_{uu}+L^+_{vv}-2L^+_{uv}}$ where L^+ is the Pseudoinverse Laplacian Matrix	$O(v^3)$	Liu and Lu, 2010

Quasi-local approach

Quasi-local similarity-based measures balance performance and complexity. Some of these metrics achieve efficiency levels comparable to those of local approaches. And also include additional topological information, as the global methods do (Martinez, et al. 2017).

Table 8: Quasi-local similarity-based measures for homogeneous networks

Similarity measure	Equation	Time complexity	Reference
Local path (LP)	$S_{LP} = A^2 + \alpha A^3$	$O(lv^2k)$	Lu et al., 2009
Shortest path (SP)	$S_{SP} = - \lvert shortest\ path(u,v) \rvert$	$O(nelogn)$	
Local Random Walk (LRW)	$S_{LRW}(u,v) = \frac{\lvert\Gamma(u)\rvert}{2\lvert E\rvert} p_{uv}(t) + \frac{\lvert\Gamma(u)\rvert}{2\lvert E\rvert} p_{uv}(t)$	$O(lv^2k)$	Liu and Lu, 2010

5.1.2.2.Bipartite networks

A bipartite network can be defined as $G = \langle V_1, V_2, E_{12} \rangle$ where V_1 and V_2 are two sets of nodes o different types, and E_{12} is the set of links between V_1 and V_2.

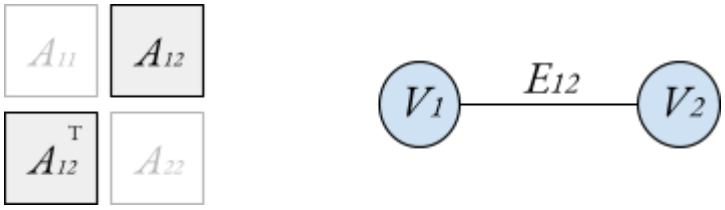


Figure 9: Block matrix and network diagram representation of a generic bipartite network

In the particular case of attributed networks, the attribute matrix X_{AT} can be interpreted as the adjacency matrix of a bipartite network between nodes $v \in V$ and topic attributes $t \in T$. Then, the topic similarity of a pair of nodes can be calculated using topology-based measures for bipartite networks. Local measures are particularly useful for this purpose since every pair of nodes is not directly connected and the local neighborhood of each node represents their relations with topical attributes that represent the nodes' interests.

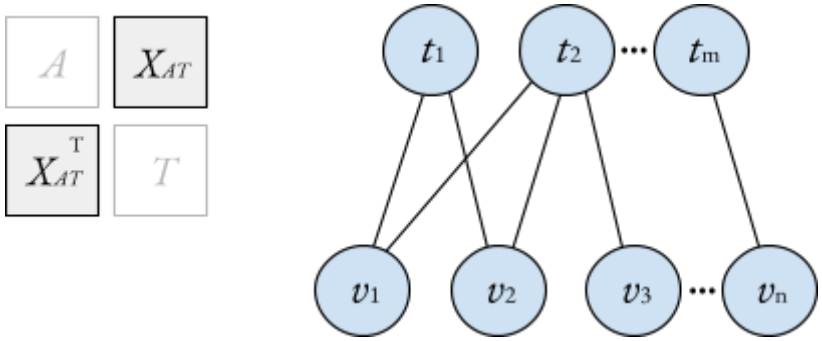


Figure 10: Block matrix and network representation of an author-topic bipartite network

Many studies that use this approach have leverage existing similarity measures like Common Neighbors (CN), Jaccard Coefficient (JA), Adamic-Adar (AA) and Resource Allocation (RA) and propose new ones (Hasan et al., 2006; Wohlfarth and Ichise, 2008; Romero et al., 2013).

Keyword-based Common Neighbors (KCN): This simple and intuitive measure is the bipartite equivalent of Common Neighbors (CN). For two nodes, u and v in V, the KCN is defined as the number of words in common between the documents associated with u and v

$$S_{KCN}(u, v) = |\Gamma(u) \cap \Gamma(v)|$$

Hasan et al. (2006) named this measure as Keyword Match Count (KMC) and applied it to a link prediction task in a co-authorship network. In this context, the measure represents the proximity of a pair of authors based on the number of shared keywords in their papers. According to the authors the larger the size of the intersection, the more likely they are to work in related areas and hence a better candidate to be a future coauthor pair. The results obtained by Hasan et al. conclude that KMC was the top ranked attribute of a set of nine features. Zhang, Shen and Wu (2019) use the same metric as temporal sequence to train a forecasting model for temporal link prediction.

Keyword-based Jaccard Coefficient (KJC): Wohlfarth and Ichise (2008) also called this coefficient as Keywords Match Count (KMC) and applied it to measure the similarity of pairs of authors using keywords extracted from the titles of their previous papers. This similarity function is defined as

$$S_{KJC}(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

Since the measure normalizes the size of the common keywords by the combined number of words, the coefficient is sensitive to small size sets. For example, Rahmaida et al. (2019) measured the Jaccard Coefficient between authors in the field of taxonomy using the occurrence in their past papers of a narrow set of research keywords corresponding to seven categories of kingdom. The results of the distributions of the coefficient show no difference between pairs of authors who initiated co-authorship and those who didn't.

When using titles as a text source, the inclusion of this measure in a supervised learning model slightly improves the performance in comparison with the study baseline (Wohlfarth and Ichise, 2008; Rahmaida et al., 2019). Sachan and Ichise (2010) reported a 15% improvement in the performance when including the Jaccard Coefficient using the abstracts of papers.

Keyword-based Adamic-Adar (KAA): Romero et al. (2013) build a bipartite network that represents the usage of keywords in a micro-blogging online social network (e.g., Twitter) and applied this measure to predict directed and mutual follow relations between users. The authors focused on a specific type of keyword called hashtag used to label conversation topics. The inverse logarithm implies that more popular common hashtags are less likely to contribute to the formation of a new link.

$$S_{KAA}(u,v) = \sum_{t \in \Gamma(u) \cap \Gamma(v)} \frac{1}{log|\Gamma(t)|}$$

Keyword-based Resource Allocation (KRA): In the same scenario as for KAA, Romero et al. (2013) also applied the Resource Allocation measure. Both measure KAA and KRA work similarly and the main difference between them is the way that they penalize more popular hashtags. While KAA uses a logarithmic function, KAA does it linearly.

$$S_{KRA}(u,v) = \sum_{t \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(t)|}$$

Smallest degree of common keywords (SDCK): Proposed by Romero et al. (2013) this measure captures the extent to which the conversation two persons share are unique or not. When analyzing the distribution of the proposed feature they found that the probability of a link as a function of the smallest common hashtag degree appears to obey an inverse power law.

$$S_{SDCK}(u,v) = \frac{1}{min_{t \in \Gamma(u) \cap \Gamma(v)} |\Gamma(t)|}$$

The authors compared this measure with KAA and KAA and found it very similar with the advantage that SDCK is cheaper to compute.

Keyword-based Hub-Promoted (KHP): Elkabani and Khachfeh (2014) proposed to use this feature to predict formation of new links in online social networks (e.g., Facebook). The measure is defined as as the number of common keywords divided by the minimum size of the two sets of keywords. The keywords represent the users' attributes and interests. The attributes selected by the authors for use in prediction are activities athletes, education, games, interests, languages, movies, music, teams, TV shows, work, books, groups, and page likes.

$$S_{KHP}(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{min(|\Gamma(u)|, |\Gamma(v)|)}$$

The authors found that for the same selection of features the keyword-based hub-promoted similarity measure provided better results than that of the Jaccard coefficient.

5.1.2.3.Semi-bipartite networks

In contrast to bipartite networks, in a semi-bipartite network links between nodes of one of the sets of nodes are allowed. This can be expressed as $G = \langle V_1, V_2, E_{12}, E_{22} \rangle$ where E_{12} is the set of links between the groups of nodes V_1 and V_2, whereas E_{22} depict the links among the nodes in V_2. In this configuration V_2 can be considered as the center group.

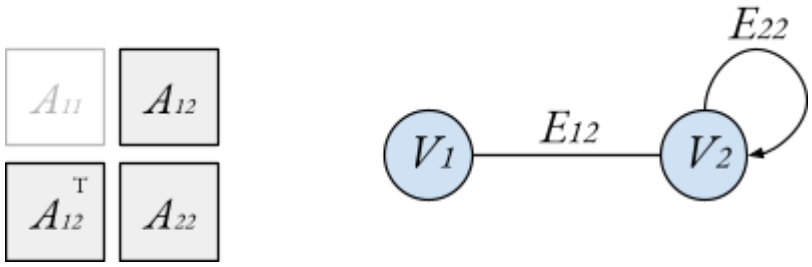


Figure 11: Block matrix and network diagram representation of a generic semi-bipartite network

Shortest path (SP): Hasan et al. (2006) apply this well-known topological measure in an author-keyword semi-bipartite network, where keywords are the center group. Therefore, two keywords that appear together in any paper are connected by a link. This measure is found to be the lowest ranked attribute among a list of 9 features. In contrast, the shortest distance in the author-author network is the top ranked among the topological features.

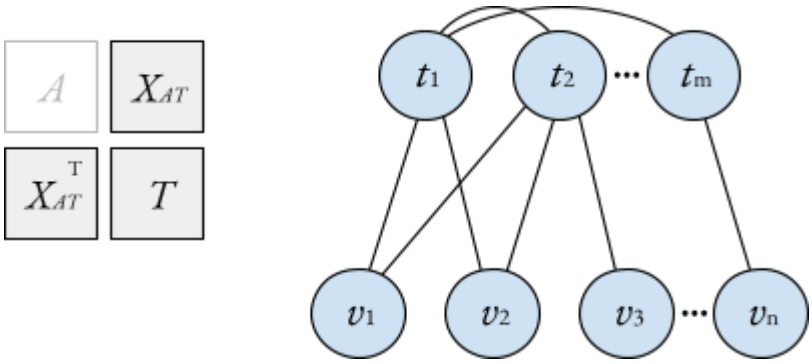


Figure 12: Block matrix and network representation of an author-topic semi-bipartite network where authors are the center group.

Katz score (Katz): uses the Katz score to estimate the similarity between authors in an author-topic semi-bipartite network. The author-word matrix X_{AW} is modeled using LDA obtaining

the author-topic matrix X_{AT}, then the topic-topic T matrix is obtained as the product of $X_{AT}{}^{t}X_{AT}$. By composing X_{AT} and T matrices in a block matrix Y, the Katz Score is computed for every pair of nodes generating a similarity author-author matrix. Then the Katz scores are converted to a binary adjacency matrix A by estimating a threshold.

$$S_{Katz}(u,v) = \sum_{L=1}^{\infty} \alpha^{L} \left| path_{L}(u,v) \right|$$

In matrix notation, the similarity can be expressed as

$$S_{Katz}(Y) = \alpha Y + \alpha^{2}Y^{2} + \cdots = (I - \alpha Y)^{-1} - I$$

When comparing Katz score to node-based similarity measures Makrehchi (2011) found that the cosine similarity based on the bag-of-words representation works very well for recommending single link to every node, but for predicting the whole structure of the network Katz's measure outperforms the node-base approach.

Forest model weak similarity (FMWS): Bhattacharyya et al. (2011) proposed a Forest Model representation of the topic-topic matrix T, where topics correspond to keywords denoting users interests in an online social network. The Forest Model is generated based on the semantic relationships of words from a large lexical database named WordNet (Fellbaum 1998). The forest generation heuristics consider relations like hypernyms and hyponyms, holonyms and meronyms, and synonyms and similars. The result it's a network composed of multiple hierarchical trees of related words. Then the weak similarity measure is defined as

$$S_{FMWS}(u,v) = \frac{\sum_{k_i \in \Gamma(u),\, k_j \in \Gamma(v)} n(k_i,k_j)}{|\Gamma(u)| \cdot |\Gamma(v)|}$$

Where $n(k_i, k_j)$ equals 1 if both keywords belong to the same tree and 0 in the other case.

$$n(k_i,k_j) = \begin{cases} 1, & d(k_i,k_j) \neq \infty \\ 0, & d(k_i,k_j) = \infty \end{cases}$$

Where $d(k_i, k_j)$ is the distance between nodes defined as the sum of the distance between both keywords i and j and their least common ancestor. This definition is equivalent to the length of the shortest path. In the case that two keywords belong to unconnected trees the distance between them is considered as infinite.

Forest model strong similarity (FMSS): The weak similarity proposed by Bhattacharyya et al. (2011) simplifies the relationships between keywords into binary values. In contrast, the strong similarity approach leverages the negative exponential function to transform the distance values to a continuous variable between 0 and 1, preserving the same values in boundary conditions. The strong similarity is defined as.

$$S_{FMSS}(u,v) = \frac{\sum_{k_i \in \Gamma(u),\, k_j \in \Gamma(v)} e^{-d(k_i,k_j)}}{|\Gamma(u)| \cdot |\Gamma(v)|}$$

PageRank Preferential Attachment (PRPA): Nigam and Chawla (2016) applied the semi-bipartite framework to the task of topic recommendation, where the relations of interest are the new author-topic links. The study compares traditional topology-based measures like Common Neighbors (CN), Jaccard Coefficient (JC), Adamic-Adar (AA), and Preferential Attachment (PA).

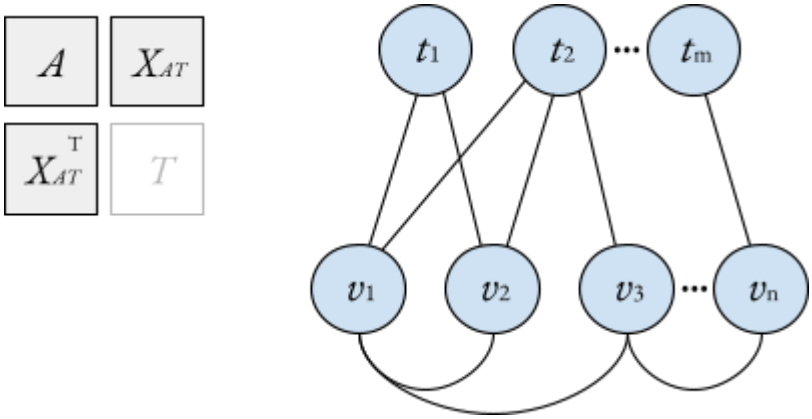


Figure 13: Block matrix and network representation of an author-topic semi-bipartite network where topics are the center group.

In addition, they proposed a measure based on the well known PageRank centrality (Page et al., 1999). This measure estimates the similarity between two nodes u and v as the product of the nodes' respective PageRank centrality. In matrix notation PageRank centrality can be expressed as following

$$PR = D(D - \alpha Y)^{-1} \cdot \beta$$

Where D is the diagonal matrix with elements $D_{ii} = max(|\Gamma(i)|, 1)$, and β is a vector of ones. Then the similarity is calculated as the product of the centrality of both nodes:

$$S_{PRPA}(u,v) = PR(u) \cdot PR(v)$$

Although the authors do not discuss about the performance of the proposed similarity measure this could be easily adapted to address the link prediction task

5.1.2.4.Heterogeneous networks

Given a set of nodes $V = \{V_i\} : i \in [1, h]$ where V_i is the subset of nodes of type i, and a set of links $E = \{E_{ij}\} : i \in [1, h], j \in [1, h]$, the graph $G = \langle V, E \rangle$ is called an heterogeneous network. This type of network can generate different configurations depending on the information available.

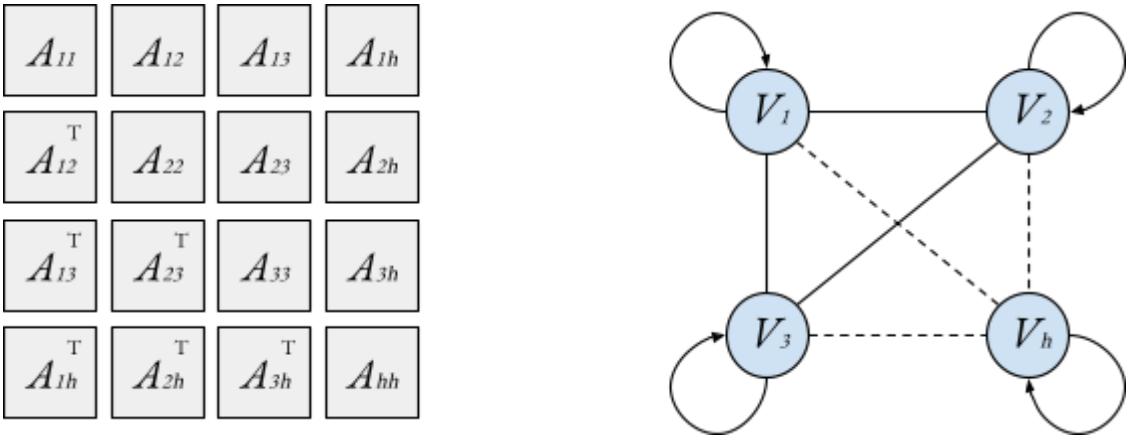


Figure 14: Block matrix and network diagram representation of a generic heterogeneous network

Confidence Index (CI): Huang, Contractor and Yao (2009) proposed a knowledge-based recommendation system for social networks. The text data is represented as keywords in a three-layer network between authors, papers, and keywords. In this network configuration only the authors are allowed to link with nodes of the same type.

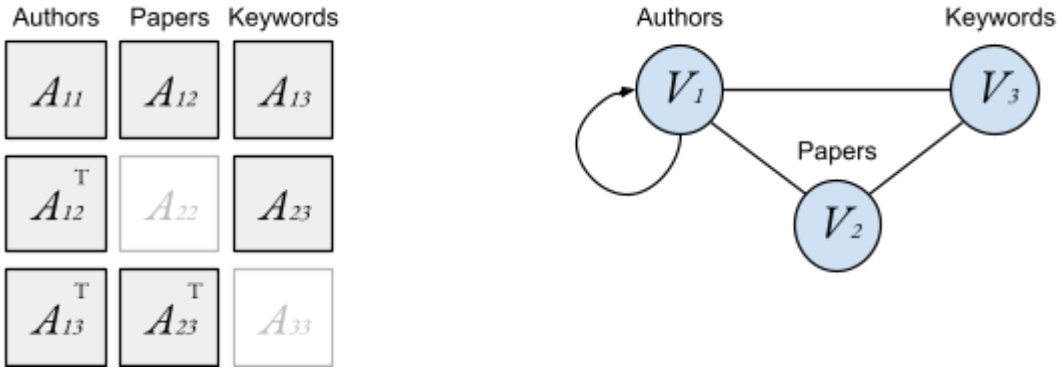


Figure 15: Block matrix and network diagram representation of an author-keyword-paper heterogeneous network.

The CI similarity is the result combining four similarity measures. The first three, measure the shortest path G_{kv}, positive match count P_{kv}, and euclidean distance D_{kv} between the search keyword k and the recommended item v. Meanwhile the fourth, measures the reciprocal of the shortest path $1/G_{uv}$ between the author u and the recommended item v.

$$S_{CI}(u,v) = \frac{[G_{max} - G_{kv}] + [P_{kv}/P_{max}] + [(D_{max} - D_{kv})/100 * D_{max}] + [1/G_{uv}]}{G_{max} + 1 + 0.01 + 1}$$

Where G_{max} is the network diameter, P_{max} is the maximum match count, and D_{max} the maximum euclidean distance in the network. Although this measure is not specifically designed for the link prediction task it can be adapted for this purpose by imposing some restrictions to the recommendation.

Meta Path Similarity Framework: Sun et al. (2011) proposed a path-based framework for modeling relations in heterogeneous networks with a star network schema (Sun et al., 2009). In this configuration nodes can only interact with nodes of a central type. The main application is bibliographic information networks formed by authors, papers, keywords and venues, where papers are the central type that connects with the other entities. The star schema also allows citation relations between papers. In contrast to the other entities, the keywords are derived from the papers' title extracting frequent phrases from the text.

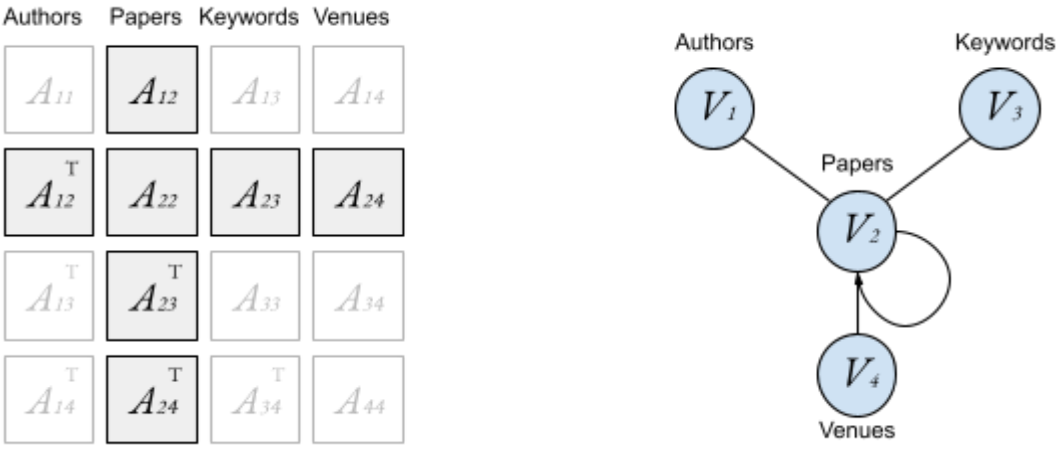


Figure 16: Block matrix and network diagram representation of an author-keyword-paper-venues heterogeneous network.

The framework defines a meta path as the path from one node to another across the different types of nodes. For example, the relation of two authors that share a common neighbor can be expressed as the meta path author-paper-author-paper-author or $APAPA$. The meta paths

$APKPA$ and $APVPA$ represent two authors that have published about the same keyword or in the same venue respectively.

Path count (PC):

The path count is the number of path instances between two nodes following a given meta path. The measure is calculated as the product of the adjacency matrices associated with the relation R represented by the meta path.

$$S_{PC}^{R}(u,v) = \prod_{i=1}^{n-1} A_{r_i r_{i+1}}(u,v)$$

Let the list $r = (1,2,1,2,1)$ denote the node type index associated with the meta path $R = APAPA$, the path count matrix is calculated as $S_{PC}^{APAPA} = A_{12}A_{21}A_{21}{}^{T}A_{12}{}^{T}$.

Normalized path count (NPC):

In the same fashion as the Sørensen Index the NPC weighs the number of paths connecting two authors by their overall connectivity (Sun et al., 2011). Normalizing the path count makes NPC less sensitive to outliers due to highly connected nodes.

$$S_{NPC}^{R}(u,v) = \frac{2 \cdot S_{PC}^{R}(u,v)}{S_{PC}^{R}(u,\cdot) + S_{PC}^{R}(\cdot,v)}$$

Path-based Random Walk (PRW):

It is the probability of the random walk that starts from u and ends with v following meta path R (Sun et al., 2011). The measure is defined as the sum of the probabilities of all the path instances which can be expressed as.

$$S_{PRW}^{R}(u,v) = \frac{S_{PC}^{R}(u,v)}{S_{PC}^{R}(u,\cdot)}$$

***Path-based Symmetric Random Walk (PSRW)*:**

The similarity measure considers the random walk from two directions along the meta path (Sun et al. 2011).

$$S_{PSRW}^{R}(u,v) = S_{PRW}^{R}(u,v) + S_{PRW}^{R}(v,u) = \frac{S_{PC}^{R}(u,v)}{S_{PC}^{R}(u,\cdot)} + \frac{S_{PC}^{R}(u,v)}{S_{PC}^{R}(v,\cdot)}$$

Zhang (2017) compared the four similarity measures for five different metapaths (APAPA, APAPAPA, APJPA, APKPA, APKPKPA). The NPC measure was the best rated feature across all five metapaths. In addition, using a Logistic Regression model different combinations of the five metapath features were compared. The results show that the combination of all five metapaths

Implicit User Interest Similarity (IUIS): Zarrinkalam et al. (2016) proposed a tweet recommendation application for Twitter's users by inferring implicit topical interests. This task is analogous to the link prediction problem however the predictive approach is focused on the definition of meaningful attributes rather than on a novel similarity metric.

The authors represented social and semantic relations as an heterogeneous network composed by three subgraphs user-user A_{11}, user-topic A_{14} and topic-topic A_{44} derived from the tweets corpus. In this framework topics represent non-overlapped communities of concepts detected from a concept correlation network A_{33} using the Louvain algorithm (Fani et al., 2015).

Regarding the definition of topic relatedness in the topic-topic network A_{44} the authors proposed three alterative approaches, namely: wikipedia-based semantic relatedness, user-based collaborative filtering relatedness, and hybrid relatedness that combines both approaches.

Finally, the prediction task is tested using traditional topology-based methods: Adamic-Adar, Common Neighbors, Jaccard Coefficient, Katz, and SimRank. The best performance is achieved by combining the semantic relatedness approach with the Adamic-Adar similarity.

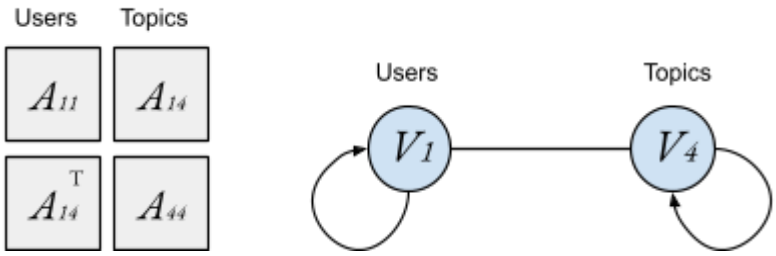


Figure 17: Block matrix and network diagram representation of an author-topic heterogeneous network.

Explicit User Interest Similarity: Similarly to Zarrinkalam et al. (2016) Amiri and Shobi (2017) present a content recommendation system for Twitter's users based on their interests and social connections. However, Amiri and Shobi (2017) proposed an explicit approach that differentiates from the implicit one in that it focuses on the user shared content (e.g., tweets, and retweets) rather than on the user interests represented as latent topics.

In this approach, the tweets' text is represented using the Doc2vec embedding model. The output is a space vector model where individual tweets are represented by a K-dimensional vector. For practical reasons, we refer to dimensions in the model as topics. The relationship between tweets are computed as the cosine similarity between their respective topic vectors.

Once the heterogeneous network is completely built the link prediction task is tested using the Jaccard Coefficient.

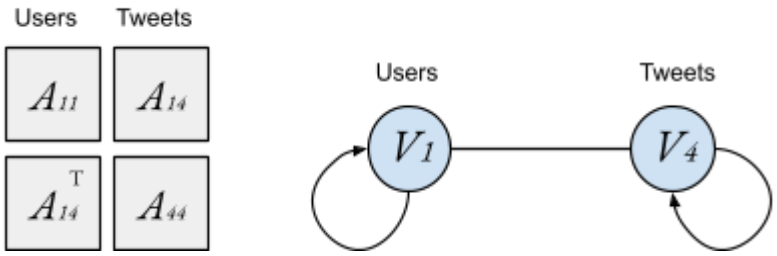


Figure 18: Block matrix and network diagram representation of an author-tweet heterogeneous network.

5.1.3. Hybrid similarity measures

Profile Similarity (PSIM): Armenato and Godoy (2013) proposed a user recommendation strategy for finding new friends to follow in micro-blogging online services like Twitter. The strategy consists of two steps. First, a candidate list R is generated from the directed network topology as:

$$R = P - K$$

$$K = \Gamma_{out}(u)$$

$$L = \cup_{k \in K}\{\Gamma_{in}(k)\}$$

$$P = \cup_{l \in L}\{\Gamma_{out}(l)\}$$

Where K is the outward directed neighborhood of u and denotes the set of users followed by the target user, L is set of the followers of all users in K, and P represents the set of users followed by the all users in L.

Second, the text-based similarity between the target user u and each candidate $v \in R$ in the list is computed. The authors proposed three similarity measures considering the attribute matrix X as the word frequency in the users' generated posts.

Profile Similarity 0: Measures the direct similarity between the users' profiles

$S_{PSIM0}(u, v) = S_{COS}(x_u, x_v)$

Profile Similarity 1: Measures the similarity between the candidate user and the average profile of the users followed by the target user.

$S_{PSIM1}(u, v) = S_{COS}(x_u^{avg}, x_v)$ where $x_{u,i}^{avg} = \frac{1}{|\Gamma_{out}(u)|} \cdot \sum_{z \in \Gamma_{out}(u)} x_{z,i}$

Profile Similarity 2: Measures the highest similarity between the candidate user and any of the users followed by the target user.

$S_{PSIM2}(u, v) = max(\{S_{COS}(x_z, x_v) : z \in \Gamma_{out}(u)\})$

Although these measures are proposed for link recommendation in directed networks they can be easily adapted for link prediction in undirected networks by taking the mutual similarity.

Contents and Collaboration Networks for Collaborators Recommendation (CCREC): Kong et al. (2016) proposed an hybrid algorithm based on word2vec representation of text and the topology-based algorithm Random Walk with Restart (RWR). The link prediction process starts with the extraction of the word co occurrence matrix W from the titles of the corpus' documents. Then W is used to train a word2vec model and the word-topic matrix X_{WT} is computed by clustering the obtained vectors into K topics by using the K-means algorithms. The X_{WT} matrix is then used to compute the author-topic matrix X_{AT}. This matrix represents the interest of authors on each research topic

The co-authorship matrix A is partitioned generating K different networks, one for each topic. Using the RWR algorithm the rank value for each author in each domain is calculated. The resulting matrix SI corresponds to the rank scores of RWR for each author on each topic, and represents the strength of influence of a researcher on a given research domain.

Finally the similarity between two authors u and v, is then calculated using the Cosine Similarity (COS) based on their feature vectors $SI(u)$ and $SI(v)$.

$S_{CCREC}(u, v) = S_{COS}(SI(u), SI(v))$

LSA based Cosine Similarity (LDACOSIN): Chuan et al. (2017) proposed a similarity measure that combines text processing and the local topology of nodes for co-authorship networks. In this type of network two nodes are connected if they have published one or more papers together. The topic distribution for each paper x_p is estimated using topic modeling method LDA.

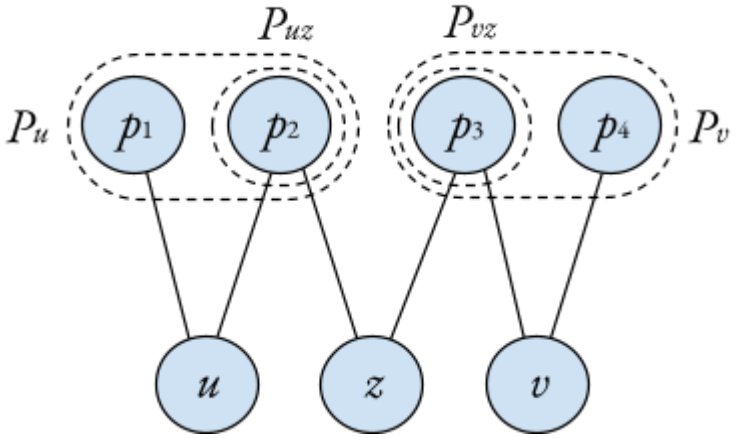


Figure 19: Bipartite network diagram illustrating the set of papers P_u , P_v , P_{uz} and P_{vz}.

The similarity measure is calculated combining two terms. The first term $S(P_u, P_v)$ is the similarity degree between two sets of papers P_u and P_v written by the authors u and v respectively. The second term is the average among all common neighbors z between u and v of the similarity degree between P_{uz} and P_{vz} that represent the sets of papers written by the pairs of authors u, z and v, z correspondingly.

$$S_{LDAcosin}(u,v) = S(P_u, P_v) \cdot \frac{1}{|\Gamma(u) \cap \Gamma(v)|} \cdot \sum_{z \in \Gamma(u) \cap \Gamma(v)} S(P_{uz}, P_{vz})$$

Each set of papers P_u, P_v, P_{uz} and P_{vz} is represented by a K-dimension vector x_u, x_v, x_{uz} and x_{vz} respectively that is computed by taking the average value for each topic among the topic distributions of the papers in the set.

$$x_{u,i} = \frac{1}{|P_u|} \sum_{p \in P_u} x_{p,i}, \; i = [1, K]$$

And the similarity degree between two sets of papers is given by

$$S(P_u, P_v) = \frac{1}{e^{1 - COS(x_u, x_v)}}$$

Structure Topics Prediction (STP): Proposed by Solaimannezhad and Fatemi (2017) this similarity measure is the direct product between the cosine similarity between the node attributes and the Adamic-Adar topology-based similarity measure. Similarly to LDACOSIN (Chuan et al. 2017) this measure aims to predict new scientific collaboration in a co-authorship network by

modeling a corpus of academic abstracts using LDA. The attribute vector x_u of an author u is a K-dimension vector extracted from his authorship.

$$S_{STP}(u,v) = S_{COS}(u,v) \cdot S_{AA}(u,v) = \frac{x_u \cdot x_v}{||x_u|| \cdot ||x_v||} \cdot \sum_{z \in \Gamma(u) \cap \Gamma(v)} \cdot \frac{1}{log|\Gamma(z)|}$$

Network structure and topic distribution (NTSD) indices: It is a set of indices proposed by Huang et al. (2020) to train different off the shelf classification models to perform the link prediction task in a directed social network. Here we list some of the proposed indices.

Homophily index (S_{HI}): Corresponds to the Pearson correlation coefficient S_{PC} between the users' topic distributions obtained from a LDA model.

$$S_{HI}(u,v) = S_{PC}(u,v)$$

Transitivity indices (S_{T2}, S_{T3}): Corresponds to the mean and variance of topic entropies of common neighbors. The topic entropy $E(z)$ is defined as Shannon information entropy of users' topic distributions.

$$E(z) = -\sum_{i=1}^{m} x_{z,i} log\, x_{z,i}$$

$$S_{T2}(u,v) = \frac{1}{|\Gamma(u) \cap \Gamma(v)|} \cdot \sum_{z \in \Gamma(u) \cap \Gamma(v)} E(z)$$

$$S_{T3}(u,v) = \frac{1}{|\Gamma(u) \cap \Gamma(v)| - 1} \cdot \sum_{z \in \Gamma(u) \cap \Gamma(v)} \left(E(z) - T_2\right)^2$$

Clustering indices: The mean and the maximum of Homophily index between the user v and the neighbors of the user u.

$$S_{C1}(u,v) = \frac{1}{|\Gamma(u)|} \sum_{z \in \Gamma(u)} S_{HI}(v,z)$$

$$S_{C2}(u,v) = max(\{S_{HI}(v,z): z \in \Gamma(u)\})$$

Degree-heterogeneity indices: Mean and variance of topic entropies of neighbors of user u

$$S_{DH1}(u) = \frac{1}{|\Gamma(u)|} \cdot \sum_{z \in \Gamma(u)} E(z)$$

$$S_{DH2}(u) = \frac{1}{|\Gamma(u)|-1} \cdot \sum_{z \in \Gamma(u)} \left(E(z) - D_1\right)^2$$

These indices were used to train different learning-based classification models such as Logistic regression (LR), Support vector machine (SVM), and Random forest (RF). The RF model performed the best, however the individual performance or contribution of the individual indices was not reported.

5.2.Learning based models

Existing learning-based models for text-based link prediction, comprises a wide range of methods that can be categorized into three groups: probabilistic and statistical models, graph embedding and dimensional reduction, and feature-based classification models. These categories can be traced to broader taxonomies for general-purpose link prediction methods and graph-embedding techniques (Martínez et al., 2017; Cai et al., 2017; Yuliansyah et al., 2020).

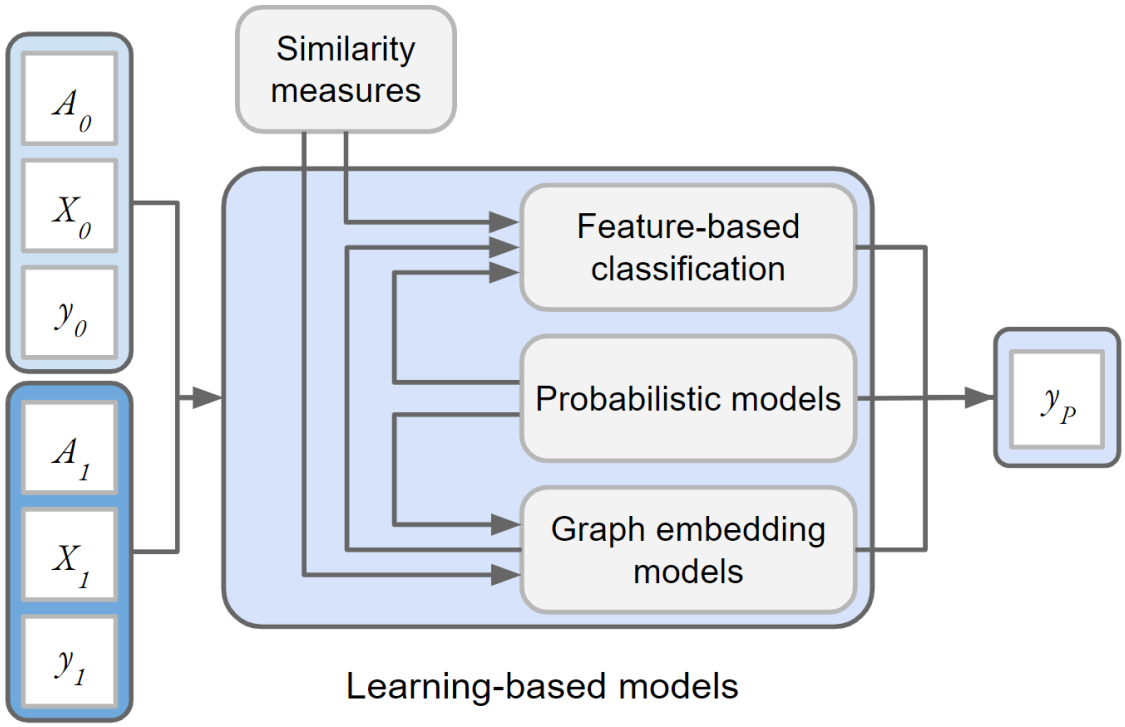


Figure 20: Learning-based models for link prediction

It is important to note that many of the text-based link prediction techniques are built incrementally by leveraging existing methods as building blocks. This leads to more complex relationships between categories than is often perceived (see Figure 20). Particularly when new techniques are usually presented as developments that start from scratch. For example, it is common to find feature-based classification models trained using combinations of similarity-based measures (Hasan et al., 2006; Zhang, 2017), as well as node attributes derived from probabilistic topic models (Liu et al., 2009), and graph embedding models (Gao et al., 2018; Ahmed et al., 2020; Berahmand et al., 2021). Similarly, there are cases of graph embedding models that use text-based similarity measures (Ahmed et al., 2020) and node attributes derived from probabilistic topic modes (Pham and Do, 2021; Xu et al., 2021) as input.

5.2.1. Probabilistic and statistical models

The use of probabilistic models for analyzing textual data have a long history in the text mining field (Gentzkow, Kelly, and Taddy, 2019). In section 4.3 we introduced some probabilistic topic models typically used for text representation in a wide range of applications.

Here we explain how some of these methods have been applied for text-based link prediction tasks to model personal interest in topics, and the linkage probability between nodes in the network.

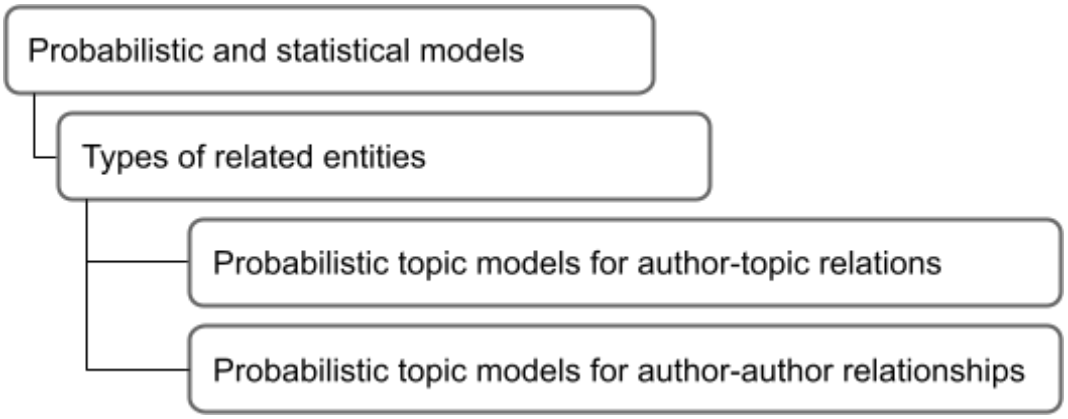


Figure 21: Taxonomy of probabilistic and statistical models

5.2.1.1. Probabilistic topic models for author-topic relations

Probabilistic topic models provide a low-dimensional representation of large collections of documents under the assumption that these documents' content can be modeled as probability distributions of latent topics. In link prediction applications, topic models have been used to estimate the nodes' attribute vectors based on the topic distributions of the documents that they are associated with.

Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan, 2003) is probably the most popular model of this kind and it has been the basis for a whole family of topic models. Previous work in link prediction has concluded that the use of LDA as complement to topology-based features increases the prediction accuracy (Makrehchi, 2011; Chuan et al., 2017; Solaimannhezad, 2017; Liu et al., 2019).

Author-Topic Model (ATM) leverages the authorship information to better estimate the authors' topic distributions (Rosen-Zvi et al., 2004). Traditional local-based link prediction methods search for potential links in two-degrees neighborhood, ATM has been used in research collaboration recommendation applications considering network distances up to four degrees (Chaiwanarom et al. 2010; Chaiwanarom and Lursinsap., 2014).

The independence of the parameters is one of Dirichlet distribution's main advantages since it decreases the dimensionality of the model. Naturally LDA assumes the independence of the estimated topics. In contrast, the Correlated Topic Model (CTM) (Blei and Lafferty, 2007) proposes that documents are generated through a multinomial process allowing correlations between the estimated topics. This approach not only provides a more realistic representation of documents but also allows to incorporate structural covariates into the model estimation and estimating their effects, for example in Structural Topic Model (STM) (Roberts et al., 2013).

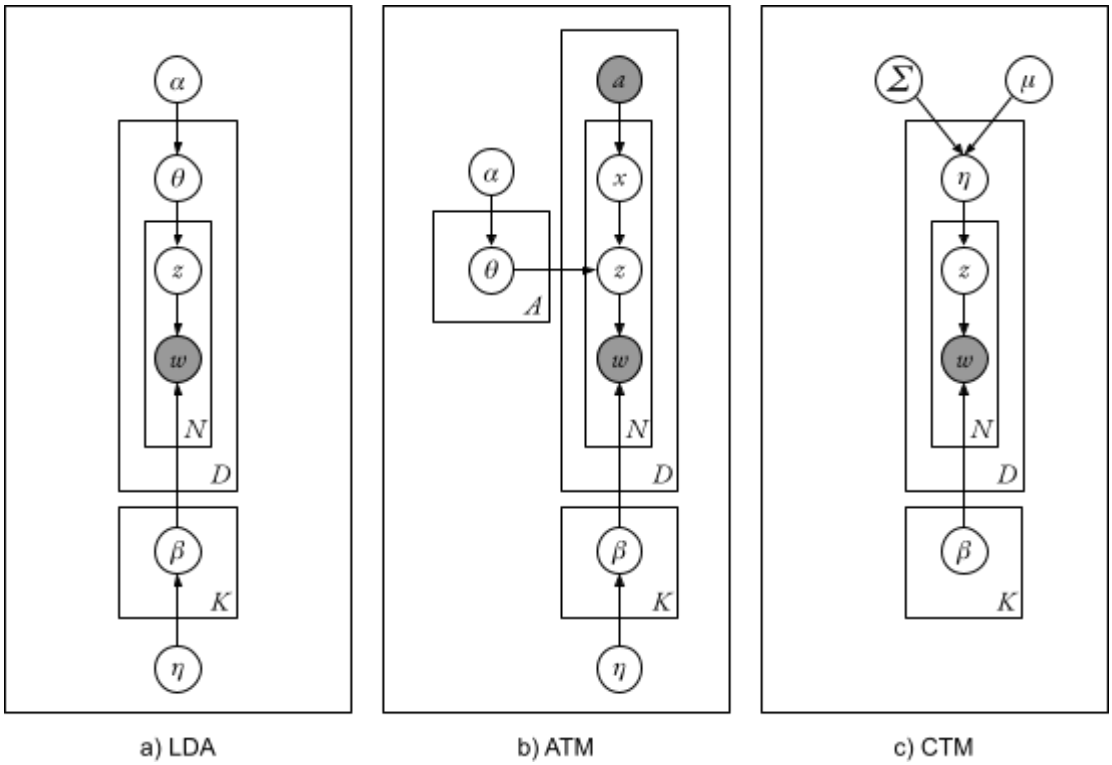


Figure 22: Plate notation diagrams probabilistic topic models LDA, ATL and CTM.

Topic models that incorporate contextual information as covariates are becoming more frequent. For example, the Supervised Latent Dirichlet Allocation model (SLDA) jointly estimates the documents' topic distributions and a response variable associated with each document (McAuliffe and Blei, 2007).

Supervised Latent Dirichlet Allocation with Covariates (SLDAX) is a generalization of SLDA, that incorporates observed variables and latent topics as predictors of an outcome by combining a latent variable measurement model and a structural regression model (Wilcox, et al. 2021).

Labeled LDA (L-LDA) constrains LDA by defining a one-to-one correspondence between LDA's latent topics and user tags, improving its expressiveness over traditional LDA (Ramage et al., 2009; Quercia et al., 2012).

In Comparative Latent Dirichlet Allocation (CompareLDA) pairwise comparison observations between documents are incorporated in the model estimation, allowing to differentiate entities along some dimension of interest (e.g., product review, movie ratings) (Tkachenko and Lawn, 2019). Semi-supervised Topic Classification model (SeedLDA) allows document classification into predefined categories efficiently by training the model with a custom “seed word dictionary” (Watanabe and Zhou, 2022).

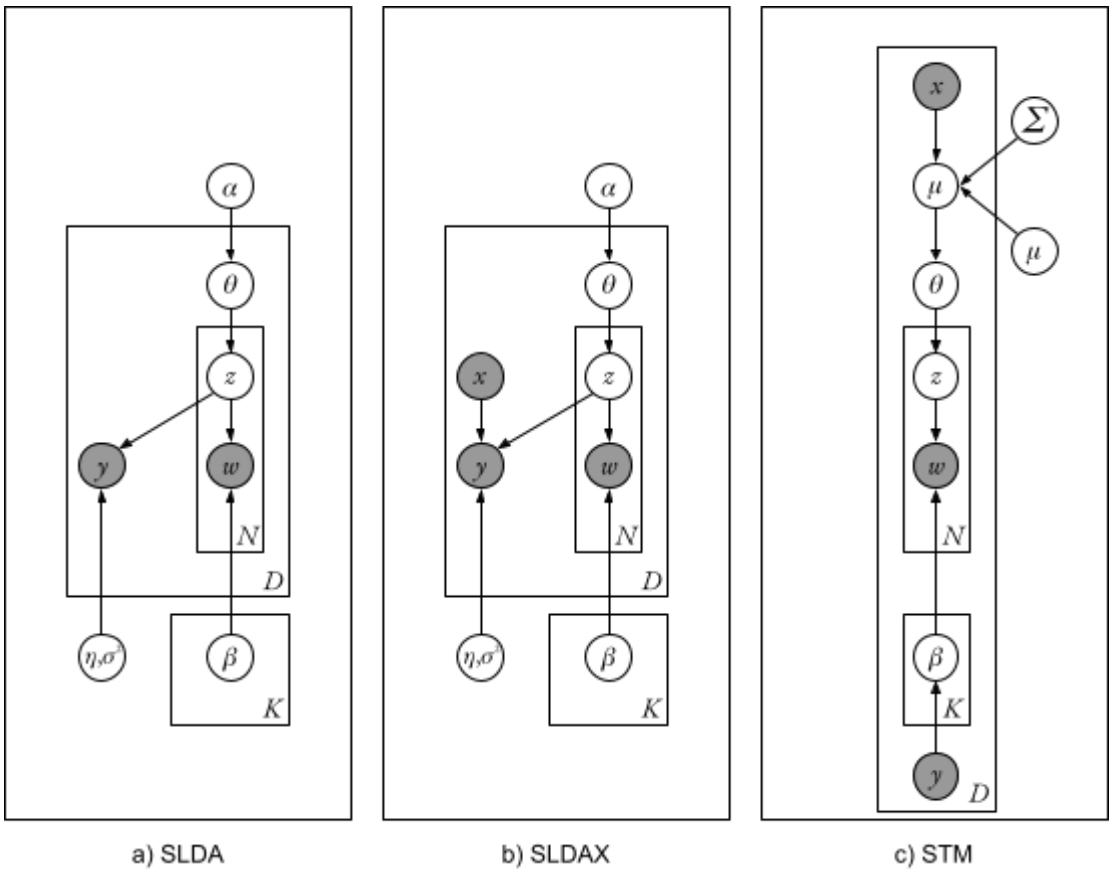


Figure 23: Plate notation diagrams supervised probabilistic topic models SLDA, SLAX and STM.

All these probabilistic topic models offer different solutions to modeling large corpus of documents whose topic distributions are useful to estimate the topical attributes of the nodes and therefore to calculate node-base similarity measures for link prediction. Depending on the context some models may have advantages over others, choosing the appropriate topic model for a specific application can significantly improve the performance of the link prediction method.

5.2.1.2. Probabilistic topic models for author-author relationships

The development of probabilistic topic models have enabled researchers to represent not only the content of large collections of documents but also the interconnections between them. Several topic models have been proposed for solving the link prediction problem in document-document networks; these models can be extended to other entities like authors (Kataria, 2012).

Mixed membership models also known as Link-LDA extend the traditional LDA model by incorporating citation links as another outcome of the generative process. This means, that the same document-specific distribution over topics θ is used to draw the latent topics for words z_n and citation links z_l respectively (Erosheva et al., 2004; Nallapati et al., 2008).

The Pairwise Link-LDA model builds upon Link-LDA but conceiving the generative process as a pairwise comparison of documents, d and d', and modeling the presence or absence of interactions (i.e., citations) between them as a binary random variable c from a Bernoulli distribution. As in Link-LDA the words and links are conditioned to different latent topics generation processes z_n and z_l. Due to these features the model is capable of estimating the whole network structure and thus providing a more general representation of the citation mechanism. However, the explicit modeling of the presence and absence of links makes it infeasible to scale to large corpora (Nallapati et al., 2008).

To address the scale problem Nallapati et al. (2008) developed the Link-PLSA-LDA by leveraging the scalability of the Link-LDA model and using Probabilistic Latent Semantic Analysis (PLSA) to explicitly model the topical dependence between cited and the citing documents. Although it is suitable for larger corpora the model assumes the citation network has a bipartite structure where each document can either be cited or be a citing document, but not both.

Similarly to Pairwise Link-LDA, the Relational Topic Model (RTM) models the link between each pair of documents as a binary random variable but RTM differentiates in that the link variable is conditioned on the same latent topics that generate the word content in the document. This allows to model the whole network structure and to give predictive distributions for links given words and words given links for new documents not present in the training set (Chang and Blei, 2009)

Arguing that a citation between two documents is not purely due to content similarity Liu et al. (2009) introduced the Topic-Link LDA model, that extends LDA and combines it with implicit author community information. In RTM the link probability function between each pair of documents is modeled as a logistic regression parameterized by coefficients η. In contrast, the generative process of links in Topic-Link LDA is conditioned to a latent community variable draw from a Dirichlet distribution with parameter κ.

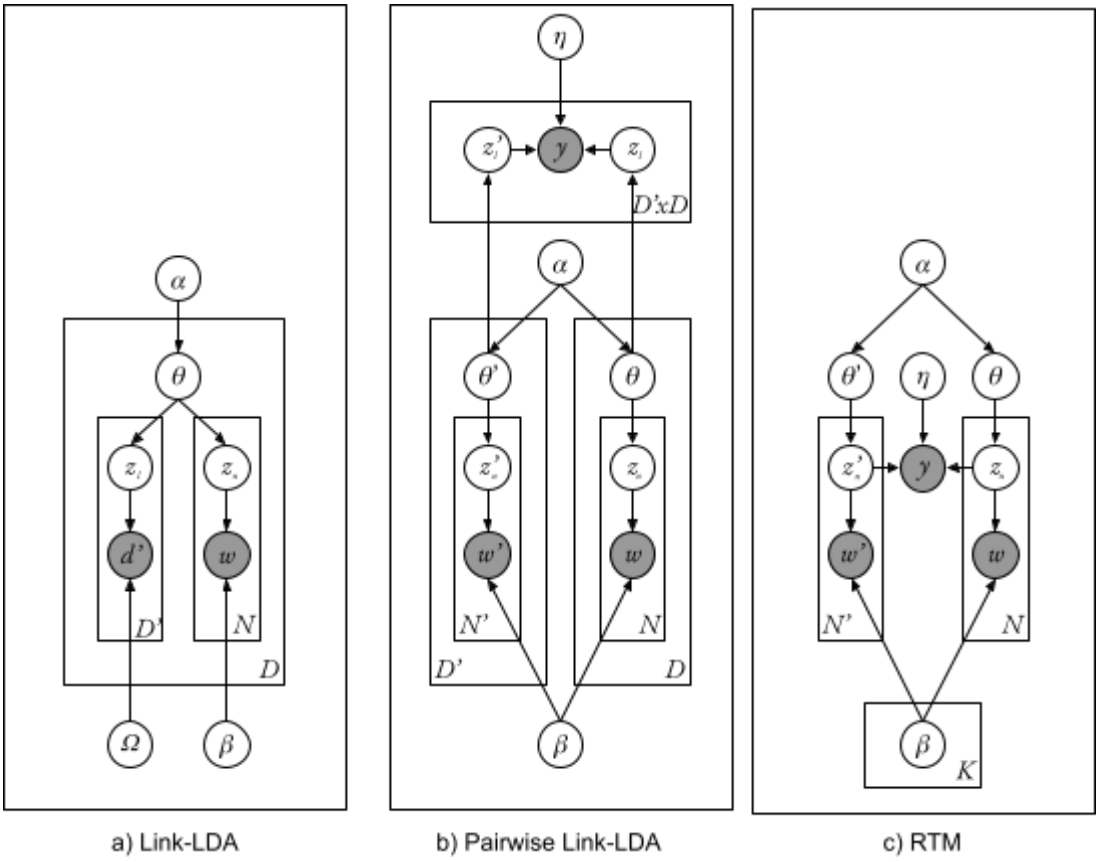


Figure 24: Plate notation diagrams supervised probabilistic topic models for topics and network Link-LDA, Pairwise Link-LDA and RTM.

Others have developed more sophisticated models for specific applications. For example, Kataria et al. (2011a) developed topic models for estimating author interests, co authorship links and author citation influence. Kataria et al. (2011b) introduced a hierarchical topic model for entity disambiguation. Tang et al. (2012) proposed the Cross-domain Topic Learning model (CTL) to address the prediction of cross-domain collaborations. Barieri et al. (2014) developed a user recommender system for social networks named “Who to Follow and Why” (WTFW) where the model decides whether the recommendation corresponds to a “topical” or “social” link. Wang et al. (2018) developed the Fusion Probabilistic Matrix Factorization Model (FMPF) which takes a LDA based similarity network and a followed/following network in a unified probabilistic matrix factorization framework.

5.2.2. Graph embedding and dimensionality reduction

The problem of graph embedding consists of representing a graph as low dimensional vectors while the graph structures are preserved (Cai et al., 2017). During the last decade several graph embedding techniques have been developed incorporating nodes attributes as labels, and text. The learned representation can be directly applied to link prediction in the form of node attributes that incarnate both structural and semantic properties.

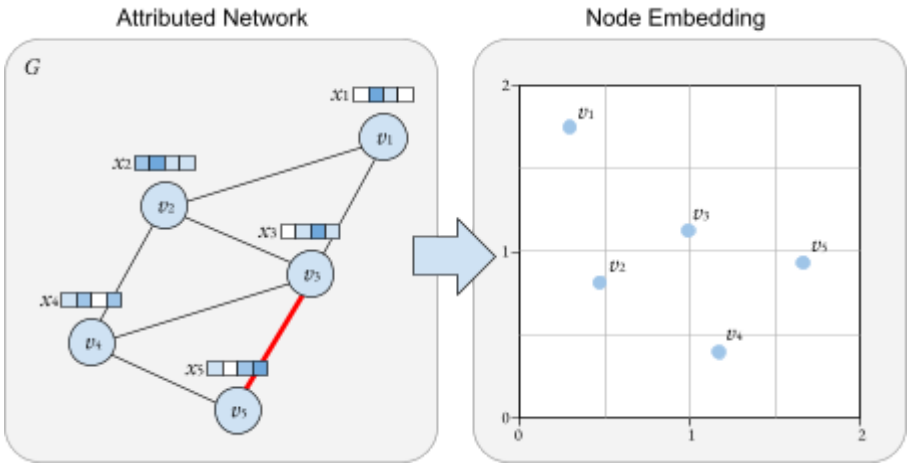


Figure 25: Representation of the node embedding process for an attributed network

Graph embedding techniques can be classified according to the type of the input network (i.e. homogeneous, bipartite and semi-bipartite, and heterogeneous networks), to the use of auxiliary information (i.e. non-attributed, attributed), and to the embedding approach (i.e. matrix factorization, deep learning with and without random walks, and others like edge reconstruction, graph kernel, generative models, and hybrid techniques) (Cai et al., 2017).

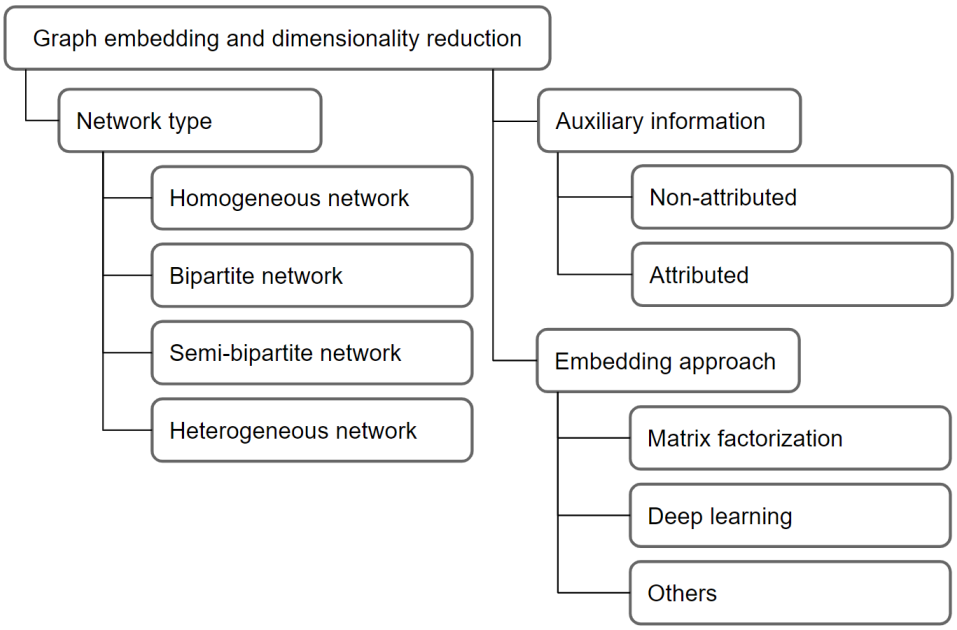


Figure 26: Taxonomy of graph embedding and dimensionality reduction models

5.2.2.1. Homogeneous networks

Most attempts on graph embedding driven link prediction are on homogeneous graphs (Cai et al., 2017). Popular methods like Deepwalk, LINE and node2vec (Perozzi et al., 2014; Tang, et al, 2015; Grover and Leskovec, 2016) have become widespread tools for analyzing homogeneous networks and predicting links. Likewise, graph embedding techniques for attributed homogeneous networks have rapidly evolved into a wide variety of methods that can be classified into three groups: Matrix factorization, deep learning with random walks, and deep learning without random walks.

Matrix factorization (MF):

Matrix factorization approaches consist in the decomposition of an input matrix $M \in \mathbb{R}^{|V|\times|V|}$ into two or more matrices, through the minimization of a loss function. The input matrix M represent the proximity of each pair of nodes in the network. There are known cases that show the ways in which text-based attributes can be integrated to the factorization algorithm (Yang et al., 2015; Huang, Li and Hu, 2017a:2017b).

In TADW (Yang et al., 2015), M is defined as $M = (A + A^2)/2$ and the decomposition task is to find matrices W and H that minimize the loss function $M - W^T HT$, where the matrix T denotes the text-based attributes of each node.

In contrast to TADW, AANE (Huang, Li and Hu, 2017a), defines the input matrix M as a combination of the adjacency matrix A and the attribute matrix T, and decompose M by minimizing $M - HH^T$ subject to $H^T H = I$, where H is the final embedding representation of the attributed network. Using a similar approach, LANE (Huang, Li and Hu, 2017b) also includes label information along with text-based attributes.

Deep learning with random walks (DL with RW):

Embedding methods based in deep learning with random walks sample the nodes' context and learn the higher order proximity between nodes. There are a number of graph embedding approaches that tackle the integration of text-based attributes into the learning algorithm (Pan et al. 2016; Hamilton, Ting and Leskovec, 2017; Liao et al, 2017; Gao and Huang, 2018; Wang, et al. 2019; Berahmand et al., 2021; Brochier and Béchet, 2021).

TriDNR model (Pan et al. 2016) builds on Deepwalk and Skipgram methods to couple two neural networks to learn the representation of the nodes' structural properties, content and labels.

GraphSAGE embedding approach (Hamilton, Ting and Leskovec, 2017) consists of first sampling the neighborhood of each node using random walks and then iteratively aggregating the attributes information from the sampled neighbors using the forward propagation algorithm.

SNE approach (Liao et al, 2017) integrates the structure and attribute modeling parts by an early fusion on the input layer. The architecture first embeds both inputs into dense vectors, and then connects them to the hidden layers. The output consists of a single low-dimensional representation consistent with both the structure and the attributes.

The DANE model (Gao and Huang, 2018) proposes a different approach which consists of two encoder-decoder branches, one for the nodes' topological structure the other for nodes' attributes. Both branches interact at the highest encoding layer through the optimization of an objective function that enforces consistency and complementarity between the low-dimensional representations of the structure and attributes.

In contrast to previous approaches, in the CSADW framework (Berahmand et al., 2021) nodes' attribute and structural data are integrated prior to sampling the random walks. The combination between the structural and attribute similarity is used as link weights to enrich the transition matrix.

Deep learning without random walks (DL without RW):

Deep learning methods can also be applied to whole graphs without the need of random walk sampling. Such methods typically combine both local network structure and node attributes applying deep neural network models with multi-layered architectures to learn low-dimensional representations of nodes (Kipf and Welling, 2016:2017, Bojchevski and Gunnemann, 2018; Meng et al., 2019; Sheikh, Kefato and Montresor, 2019).

The Graph Convolutional Networks framework (GCN) is based on a first-order approximation of spectral graph convolutions (Kipf and Welling, 2017). In this approach label attributes of nodes are seen as a multi-channel input signal which is combined with a function derived from the structure of the network through the multiplication in the Fourier domain. The output of the model is the node embeddings which can be used for estimating similarity in link prediction.

Building on GCN framework, Variational Graph Auto-Encoders (VGAE) architecture couples a GCN based inference model with Gaussian priors, and a generative model given by an inner product between latent variables, allowing to reconstruct the adjacency matrix. The learned model and reconstructed network can be used directly as a link prediction method (Kipf and Welling, 2016).

Similar to the GCN encoding used in VGAE, Graph2Gauss (G2G) method also embeds each node as Gaussian distribution allowing to capture not only the embedding parameters but also their uncertainty. The difference is that G2G combines structure and attributes vía personalized ranking instead of using approximated spectral graph convolutions. The personalized ranking approach is based on that similarity between attributes embeddings of closer nodes should rank better than more distant nodes.

Co-Embedding Attributed Networks (CAN) model, expands the VGAE framework to obtain Gaussian embeddings of both nodes and attributes by adding an attribute encoder-decoder layer to the architecture in parallel to VGAE (Meng et al., 2019).

The GLACE model stands for Gaussian representations for Large-scale Attributed graph Content Embedding, and learns node embeddings as probability distributions from both node attributes and graph structure information sequentially (Hettige et al., 2019). First, it learns the node attributes using two levels of transformations, encoding and Gaussian embedding. Then, the learned Gaussian distributions of the node attributes are used as inputs to the graph structure encoding.

The Sage2Vec model (Simple Attributed Graph Embedding) proposed an autoencoder architecture that learns the latent representation by jointly minimizing the reconstruction loss of the network structure and node attributes (Sheikh, Kefato and Montresor, 2019). In contrast to previous methods, the node attributes are considered as the objective output at the decoding stage, instead of as an input at the encoding stage.

MATAN model (Mutual Attention for Text-Attributed Networks) proposed a graph embedding procedure to train a link prediction algorithm in a network of documents (Brochier, Guille and Velcin, 2019). The method consists of learning the parameters of a mutual attention function based on the text attribute, by minimizing the reconstruction loss of the network structure.

5.2.2.2. Bipartite and semi-bipartite networks

As described in sections 5.1.2.2 and 5.1.2.3, bipartite and semi-bipartite networks model interactions between two types of nodes, therefore they can be used to represent networks with attributes, where the attributes become a second type of node. Main difference between both approaches is that bipartite networks definition only admits interaction between nodes of different types, while semi-bipartite networks additionally allow interactions between nodes of the same type.

Just like the case of homogeneous networks, deep learning methods for node embedding in bipartite and semi-bipartite graphs can be roughly classified into two groups: Deep learning with and without random walks. In addition to the embedding approach, the presence of attributes in bipartite graphs has become new classification criteria. Recent works have developed embedding techniques to deal with this kind of attributed bipartite networks by adding new information processing layers to mainstream methods (Huang, et al., 2020; Ahmed et al., 2020).

Deep learning with random walks (DL with RW):

The BINE (Bipartite Network Embedding) model learns the network representation by adopting a combined approach accounting for both explicit and implicit relations (i.e. relation between nodes of different types and nodes of the same type, respectively) (Gao et al., 2018). The explicit relations are modeled using the edge reconstruction approach proposed in the LINE model (Tang et al., 2015). While, the implicit relations are modeled in parallel by inducing two homogeneous networks, one for each type of node, and learning the node embeddings using the Deepwalk approach (Perozzi et al., 2014). Then, node embeddings are learned by jointly optimizing both objective functions.

Building on BINE's combined approach and parallel architecture, the BIGAT2VEC model is designed to allow attributes in learning both explicit and implicit relations in bipartite networks (Ahmed et al., 2020). Attributes are incorporated into the BIGAT2VEC model by inducing three pairwise similarity based networks: the explicit bipartite network and both implicit homogeneous networks. Then, All three attribute based networks are modeled in the same way as the structure based networks (Ahmed et al., 2020).

Deep learning without random walks (DL without RW):

Similar to BINE and BIGAT2VEC, BIANE (Bipartite Attributed Network Embedding) model also relies on inducing two homogeneous networks from the bipartite network, one for each type of node (Huang, et al., 2020). It differs from previous approaches by three main features: First, it uses an autoencoders architecture instead of edge reconstruction and random walks; second, the implicit and explicit relations are modeled sequentially instead of in parallel; and third, the attribute matrix is directly encoded instead of being transformed into a homogeneous network.

Intended for the detection of drug-target interaction (DTI) the SBGM-DNN model (short for Semi-bipartite Network Embedding Model - Deep Neural Network) samples positive and likely negative drug-target pairs of nodes and for each pair it induces a sub-graph that is passed as input for the deep neural network. Instead of generating low dimensional representation of nodes, the model is designed to train the deep neural network as a classifier model (Manoochehri and Nourani, 2020; Wang et al., 2021).

5.2.2.3. Heterogeneous networks

In section 5.1.2.4 we defined the concept of heterogeneous networks and how these types of network can be used to represent text-based attributes for link prediction. Recent studies in the field of graph embedding have developed methods for heterogeneous networks showing performance improvements in node classification and link prediction tasks (Dong, Chawla and Swami, 2017; Shu et al., 2021; Pham and Do, 2021; Xu et al., 2021). As with the other types of networks, graph embedding methods for heterogeneous networks can also be classified into the same categories depicted in Figure 26.

Deep learning with random walks (DL with RW):

The MetaPath2Vec model adopts the definition of metapaths in heterogeneous networks (Sun et al, 2011) and develops a framework for sampling metapaths random walks and learning the node embeddings using Skipgram (Dong, Chawla and Swami, 2017).

The AHNA (Attributed Heterogeneous Network embedding based on Aggregate-path) model leverage the nodes attributes to construct a new layer of links based on attribute proximity, then proposes a random walk strategy based on aggregate-path that adaptively chooses from which layer to sample the next node depending on the importance of node attributes and structural heterogeneity (Shu et al., 2021). Embeddings are learned using Skipgram and Bidirectional Recurrent Neural Network (BRNN) approaches.

The AHNE (Attributed Heterogeneous Network Embedding) model learns the node representation by combining three embedding approaches: autoencoders for the node attributes, LINE for the 1-step and 2-step homogeneous neighbors, and MetaPath2Vec for the heterogeneous neighbors (Wang et al., 2021).

The W-MMP2Vec (Weighted Multiple Meta-path2Vec) model is assumed as a multiple-class classification problem that predicts the existence of a metapath between a pair of nodes (Pham and Do, 2021). Text-based attributes are modeled using LDA and incorporated into the model to calculate the weight of each metapath using the average cosine similarity.

Tabla 9: Comparison of graph embedding and dimensionality reduction methods

Link prediction methods	Method	Network type	Attribute	Text processing	Approach
Perozzi et al., 2014	Deepwalk	Homogeneous	No	No	DL with RW

Grover et al., 2016	Node2vec	Homogeneous	No	No		DL with RW
Tang et al., 2015	LINE	Homogeneous	No	No		ER
Yang et al., 2015	TADW	Homogeneous	Yes	Bag-of-words	TFIDF	MF
Huang et al., 2017a	LANE	Homogeneous	Yes	Bag-of-words	Occurrence	MF
Huang et al., 2017b	AANE	Homogeneous	Yes	Bag-of-words	Occurrence	MF
Pan et al. 2016	TriDNR	Homogeneous	Yes	Word embedding	Occurrence	DL with RW
Hamilton et al., 2017	GraphSAGE	Homogeneous	Yes	Word embedding	word2vec, GloVe	DL with RW
Liao et al, 2017	SNE	Homogeneous	Yes	Bag-of-words	TFIDF	DL with RW
Gao and Huang, 2018	DANE	Homogeneous	Yes	Bag-of-words	Occurrence	DL with RW
Berahmand et al., 2021	CSADW	Homogeneous	Yes	Bag-of-words	Occurrence	DL with RW
Kipf and Welling, 2017	GCN	Homogeneous	Yes	Bag-of-words	Occurrence	DL without RW
Kipf and Welling, 2016	VGAE	Homogeneous	Yes	Bag-of-words	Occurrence	DL without RW
Bojchevski et al. 2018	G2G	Homogeneous	Yes	Bag-of-words	Occurrence	DL without RW
Meng et al., 2019	CAN	Homogeneous	Yes	Bag-of-words	Occurrence	DL without RW
Hettige et al., 2019	GLACE	Homogeneous	Yes	Bag-of-words	TFIDF	DL without RW
Brochier et al., 2019	MATAN	Homogeneous	Yes	Word embedding	GloVe	DL without RW
Sheikh et al., 2019	SAGE2Vec	Homogeneous	Yes	Bag-of-words	TFIDF	DL without RW
Gao et al., 2018	BINE	Bipartite	No	No		ER + DL RW
Huang et al., 2020	BIANE	Bipartite	Yes	Word embedding	Doc2vec	DL without RW
Ahmed et al., 2020	BIGAT2Vec	Bipartite	Yes	Bag-of-words	TFIDF	ER + DL RW
Manoochehri et al.,2020	SBGM-DNN	Semi-bipartite	No	No		DL without RW
Dong et al., 2017	MP2Vec	Heterogeneous	No	No		DL with RW
Shu et al, 2021	AHNA	Heterogeneous	Yes	Bag-of-words	Occurrence	DL with RW
Wang et al., 2021	AHNE	Heterogeneous	Yes	Bag-of-words	Occurrence	ER + DL RW
Pham and Do, 2021	WMMP2Vec	Heterogeneous	Yes	Bag-of-words	LDA	DL with RW
Xu et al., 2021	TGHNN	Heterogeneous	Yes	Bag-of-words	LDA	DL without RW

Deep learning without random walks (DL without RW):

The THGNN (Topic-aware Heterogeneous Graph Neural Network) model leverages the LDA derived text-based attributes to guide the sampling strategy by paying more attention to those metapaths higher topic consistency. Then, the metapaths are encoded into topic-aware representations and further classified in groups and aggregated depending on their inferred groups.

5.2.3. Feature-based classification

Feature-based classification algorithms are widely used in link prediction studies and applications as a complement to the main link prediction technique during the last stage of the workflow. While the link prediction technique generates proximity features for each pair of nodes, the classification algorithm learns a model and generates a binary output predicting whether a link will form or not between the nodes.

Although using classifiers may improve the link prediction performance the most critical task in this approach corresponds to the extraction and selection of an appropriate set of features for each pair of nodes (Al Hassan and Zaki, 2011). In most cases, such link prediction methods directly employ off-the-shelf classification models available in machine learning softwares like WEKA platform, LIBSVM and LIBLINEAR C/C++ libraries, or modify other well-known models (Wang et al., 2015).

Text-based link prediction techniques benefit largely from this classification approach. In general purpose link prediction methods features are commonly derived from the network structure and the nodes topological attributes (Wang et al., 2015). Text-based similarity measures contribute to the model with domain-specific context that is not contained in the topology-based features improving further the prediction performance (Hasan et al., 2006; Wohlfarth and Ichise, 2008). The development of new methods for link prediction with classifiers is in most cases achieved by constructing novel text-based similarity measures rather than by improving the existing classification models.

A key challenge of the feature-based classification approach is the issue of class imbalance. Classification models work better under optimal conditions of class balance, meaning the number of positive and negative classes should be similar. However, in large social networks the connections are sparse, and the positive cases are very rare in comparison to the negative ones. To deal with this issue pre-processing methods like oversampling the positive instances, and undersampling the negatives have shown good results (Wohlfarth and Ichise, 2008; Sachan and Ichise, 2010)

By comparing the text-based link prediction methods shown in Table 10 we observed that the use of feature-based classification is a widespread practice among similarity-based models like ones reviewed in section 5.1. Feature-based classification models can also be useful to transform probabilistic models' continuous outputs into binary predictions (Liu et al., 2009). In the case of graph embedding models, learned representations of nodes have been used as features for training a logistic regression with L2-regularization (Gao et al., 2018; Ahmed et al., 2020; Berahmand et al., 2021) and as a node attribute to compute similarity-based features for training (Pham et al., 2021).

Regarding the classification models we find that Logistic Regression (LR), Support Vector Machine (SVM) and Decision Tree (DT) are the three most used classifiers, followed by Random Forest (RF), K nearest neighbors (KNN), and Neural Network (NN).

Tabla 10: Comparison of feature-based classification methods

Link prediction methods		Features	LR	SVM	DT	RF	NB	KNN	NN	Other
Liu et al.	2009	probabilistic	LR(1)							
Gao et al.	2018	embeddings	LR_{L2}(1)							
Ahmed et al.	2020	embeddings	LR_{L2}(1)							
Berahmand et al.	2021	embeddings	LR_{L2}(1)							
Pham et al.	2021	embeddings	LR(1)							
Wang et al.	2007	similarity	LR(1)							
Sun et al.	2011	similarity	LR(1)							
Romero et al.	2013	similarity	LR(1)							
Yuan et al.	2014	similarity	LR(2)			RF(1)				
Zhang	2017	similarity	LR(1)							
Zhang and Yu	2014	similarity	LR(1)	SVM(2)			NB(4)	KNN(3)		NBM(5)
Elkabani and Khachfeh	2015	similarity	LR(4)	SVM(5)	DT(2)		NB(1)		NN(3)	
Nigam and Chawla	2016	similarity	LR(3)	SVM(1)	DT(2)					
Hassan	2019	similarity	LR(*)	SVM(*)	DT(*)	RF(1)	NB(*)	KNN(2)	NN(*)	
Huang et al.	2020	similarity	LR(3)	SVM(2)		RF(1)				
Chuan et al.	2017	similarity		SVM(1)						
Yoon et al.	2018	similarity		SVM(1)						
Zhang, Shen and Wu	2019	similarity		SVM(1)				KNN(2)		
Ho, Bui and Bui	2019	similarity		SVM(2)	DT(3)	RF(1)				
Hasan et al.	2006	similarity		SVM(1)	DT(2)		NB(5)	KNN(3)	NN(4)	
Wohlfarth and Ichise	2008	similarity			DT(1)					
Bartal et al.	2009	similarity			DT(1)				NN(2)	
Sachan and Ichise	2010	similarity			DT(1)					
Aiello et al.	2012	similarity			DT(1)					
Rahmaida et al.	2019	similarity				RF(1)				
Duricic et al.	2021	similarity								XGBoost(1)

LR: Logistic Regression; SVM: Support Vector Machine; DT: Decision Tree; RF: Random Forest; NB: Naive Bayes; KNN: K-Nearest Neighbors; NN: Neural Networks; NBM: Naive Bayes Multinomial.
*Performance evaluation was omitted for these methods.

In terms of performance, Table 10 also shows the classifiers ranking as a superscript. This ranking is constructed by sorting the models by their reported prediction performance based on the average F1-score or the area under the receiver operating characteristic curve (AUC-ROC). We observed that the RF model achieved consistently better performance than DT, SVM, and LR (Yuan et al., 2014; Hassan, 2019; Huang et al., 2020; Ho, Bui and Bui, 2019).

6.Evaluation approaches for link prediction

The assessment of link prediction models is fundamental to tell how well a model performs in the binary classification task. In most link prediction models the output set of predicted links depends on the comparison of the resulting similarity score $S(u,v)$ for each pair of nodes $\langle u,v\rangle$ to an arbitrary threshold τ.

$$E_P = \{\langle u,v\rangle \in V \times V\colon S(u,v) > \tau\}$$

Then the binary label array $y_P^{\langle u,v\rangle}$ takes value 1 if $\langle u,v\rangle$ belongs to E_P and -1 in the contrary case. Confusion matrix is common way to present the coincidences between predicted values $y_P^{\langle u,v\rangle}$ and the true values $y_1^{\langle u,v\rangle}$.

Confusion matrix		Actual values $y_1^{\langle u,v\rangle}$	
		Positive	Negative
Predicted values $y_P^{\langle u,v\rangle}$	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True Negative (TN)

where the values in the cells can be defined as:

$$TP = \left|\left\{\langle u,v\rangle\colon\ y_P^{\langle u,v\rangle} > 0 \wedge y_1^{\langle u,v\rangle} > 0\right\}\right|$$
$$FP = \left|\left\{\langle u,v\rangle\colon\ y_P^{\langle u,v\rangle} > 0 \wedge y_1^{\langle u,v\rangle} < 0\right\}\right|$$
$$FN = \left|\left\{\langle u,v\rangle\colon\ y_P^{\langle u,v\rangle} < 0 \wedge y_1^{\langle u,v\rangle} > 0\right\}\right|$$
$$TN = \left|\left\{\langle u,v\rangle\colon\ y_P^{\langle u,v\rangle} < 0 \wedge y_1^{\langle u,v\rangle} < 0\right\}\right|$$

The values in the confusion matrix are used to define several performance coefficients that measure different aspects of the model allowing the comparison with alternative prediction methods. These coefficients can be classified into two groups, fixed-threshold performance measures and curve-based measures (Yang et al., 2015).

6.1.Fixed-threshold performance measures

Corresponds to performance coefficients calculated based on the values of the confusion matrix for a given threshold τ. The coefficients change for different threshold values.

Accuracy or Rand index: Measures the number of correctly classified cases over the total number of cases. However it presents limitations in cases of class imbalance. For example, in cases where positive class is infrequent even an all negative prediction would obtain a high accuracy.

$$Accuracy = \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}$$

Recall, sensitivity or true positive rate (TPR): Corresponds to the proportion of the actual positive cases that were correctly classified. It reflects the ability of the predictor to retrieve positive cases.

$$TPR = \frac{|TP|}{|TP|+|FN|}$$

Specificity or true negative rate (TNR): Indicates the proportion of the actual negative cases that were correctly classified. It reflects the ability of the predictor to correctly exclude negative cases.

$$TNR = \frac{|TN|}{|FP|+|TN|}$$

Precision or positive predictive value (PPV): It is the proportion of positive predicted cases that were correctly guessed.

$$PPV = \frac{|TP|}{|TP|+|FP|}$$

Fall-out or false positive rate (FPR): Corresponds to the proportion of the actual negative cases that were incorrectly classified as positive.

$$FPR = \frac{|FP|}{|FP|+|TN|}$$

F1 Score: Consist of the harmonic mean between the precision and recall measures.

$$F1 = \frac{TPR + PPV}{TPR * PPV}$$

6.2. Curve-based performance measures

Threshold curves are drawn based on fixed-threshold measures by sweeping the threshold values through the whole range. Then measures are calculated by taking the area under the curves.

Area under the receiver operating characteristic curve (AUC-ROC): The ROC curve represents graphically the trade-off between the detection of positive cases (TPR) and the false positive rate (FPR). The value of the area under this curve ranges from 0 to 1. For a random predictor the AUC-ROC value is 0.5. A higher value indicates a better classification.

Area under the prediction-recall curve (AUC-PR): This curve represents the trade-off between prediction and recall. The area under the PR curve is particularly useful and informative when applied to binary classification tasks in imbalanced datasets. A higher value also indicates a better classification.

7. Conclusions

The use of text in the study of social networks is a widespread practice in the field of link prediction and related topics like community detection and node classification. In this survey we have presented, to the best of our knowledge, the most comprehensive effort to systematize the existing literature on text-based link prediction in social networks. To this end, a generic framework has been proposed that breaks down the text-based link prediction process into four well-defined phases, providing a useful tool for analyzing and designing text-based link prediction methods.

Text-based link prediction methods have emerged and have evolved into a wide variety of similarity-based and learning-based models. An adapted link prediction taxonomy is proposed to classify existing and emerging text-based link prediction methods according to their theoretical approach, and other criteria.

Learning-based methods, particularly graph embedding models, are a fast growing group in comparison to other categories. This trend is expected to continue thanks to the advance of fields like machine learning and artificial intelligence. In spite of the sophistication and complexity of the graph embedding models their use of text processing tools for attributes extraction is rather crude.

Similarity-based methods, face the opportunity of developing novel text-based attribute extraction methods that could supply learning-based methods better semantic representations of node attributes. Natural language processing tools are improving constantly posing an opportunity for future developments of text-based link prediction techniques in social networks.

8.References

Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M. (2006, April). Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security* (Vol. 30, pp. 798-805).

Wang, C., Satuluri, V., & Parthasarathy, S. (2007, October). Local probabilistic models for link prediction. In *Seventh IEEE international conference on data mining (ICDM 2007)* (pp. 322-331). IEEE.

Wohlfarth, T., & Ichise, R. (2008, November). Semantic and event-based approach for link prediction. In *International Conference on Practical Aspects of Knowledge Management* (pp. 50-61). Springer, Berlin, Heidelberg.

Bartal, A., Sasson, E., & Ravid, G. (2009, July). Predicting links in social networks using text mining and sna. In *2009 International conference on advances in social network analysis and mining* (pp. 131-136). IEEE.

Huang, Y., Contractor, N. S., & Yao, Y. (2008, May). CI-KNOW: recommendation based on social networks. In *DG. O* (pp. 375-376).

Liu, Y., Niculescu-Mizil, A., & Gryc, W. (2009, June). Topic-link LDA: joint models of topic and author community. In *proceedings of the 26th annual international conference on machine learning* (pp. 665-672).

Chaiwanarom, P., Ichise, R., & Lursinsap, C. (2010, November). Finding potential research collaborators in four degrees of separation. In the International *Conference on Advanced Data Mining and Applications* (pp. 399-410). Springer, Berlin, Heidelberg.

Chang, J., & Blei, D. (2009, April). Relational topic models for document networks. In Artificial intelligence and statistics (pp. 81-88). PMLR.

Sachan, M., & Ichise, R. (2010). Using semantic information to improve link prediction results in network datasets. *International Journal of Engineering and Technology*, 2(4), 334.

Bhattacharyya, P., Garg, A., & Wu, S. F. (2011). Analysis of user keyword similarity in online social networks. *Social network analysis and mining*, 1(3), 143-158.

Makrehchi, M. (2011, October). Social link recommendation by learning hidden topics. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 189-196).

Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C., & Han, J. (2011, July). Co-author relationship prediction in heterogeneous bibliographic networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining* (pp. 121-128). IEEE.

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2011). Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11), 992-1003.

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2), 1-33.

Quercia, D., Askham, H., & Crowcroft, J. (2012, June). Tweetlda: supervised topic classification and link prediction in twitter. In *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 247-250).

Tang, J., Wu, S., Sun, J., & Su, H. (2012, August). Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1285-1293).

Armentano, M. G., Godoy, D., & Amandi, A. A. (2013). Followee recommendation based on text analysis of micro-blogging activity. *Information systems*, 38(8), 1116-1127.

Romero, D., Tan, C., & Ugander, J. (2013). On the interplay between social and topical structure. In *Proceedings of the international AAAI conference on web and social media* (Vol. 7, No. 1, pp. 516-525).

Barbieri, N., Bonchi, F., & Manco, G. (2014, August). Who to follow and why: link prediction with explanations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1266-1275).

Elkabani, I., & Khachfeh, R. A. A. (2015). Homophily-based link prediction in the facebook online social network: a rough sets approach. *Journal of Intelligent Systems*, 24(4), 491-503.

Yuan, G., Murukannaiah, P. K., Zhang, Z., & Singh, M. P. (2014, October). Exploiting sentiment homophily for link prediction. In *Proceedings of the 8th ACM Conference on Recommender systems* (pp. 17-24).

Zhang, Q., & Yu, H. (2014). Computational approaches for predicting biomedical research collaborations. *PloS one*, 9(11), e111795.

Chaiwanarom, P., & Lursinsap, C. (2015). Collaborator recommendation in interdisciplinary computer science using degrees of collaborative forces, temporal evolution of research interest, and comparative seniority status. *Knowledge-Based Systems*, 75, 161-172.

Guo, W., Wu, S., Wang, L., & Tan, T. (2015, October). Social-relational topic model for social networks. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1731-1734).

Kong, X., Jiang, H., Yang, Z., Xu, Z., Xia, F., & Tolba, A. (2016). Exploiting publication contents and collaboration networks for collaborator recommendation. *PloS one*, 11(2), e0148492.

Nigam, A., & Chawla, N. V. (2016, March). Link prediction in a semi-bipartite network for recommendation. In *Asian Conference on Intelligent Information and Database Systems* (pp. 127-135). Springer, Berlin, Heidelberg.

Yang, W., Boyd-Graber, J., & Resnik, P. (2016, August). A discriminative topic model using document network structure. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 686-696).

Zarrinkalam, F., Fani, H., Bagheri, E., & Kahani, M. (2016, March). Inferring implicit topical interests on twitter. In *European Conference on Information Retrieval* (pp. 479-491). Springer, Cham.

Amiri, M. Z., & Shobi, A. (2017, July). A link prediction strategy for personalized tweet recommendation through doc2vec approach. In *Conference Management Team* (p. 72).

Chuan, P. M., Ali, M., Khang, T. D., & Dey, N. (2018). Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence*, 48(8), 2470-2486.

Martinčić-Ipšić, S., Močibob, E., & Perc, M. (2017). Link prediction on Twitter. *PloS one*, 12(7), e0181079.

Solaimannezhad, H., & Fatemi, O. (2017). Representing a Content-based link Prediction Algorithm in Scientific Social Networks. *Information Systems & Telecommunication*, 146.

Wang, H., Shi, X., & Yeung, D. Y. (2017, February). Relational deep learning: A deep latent variable model for link prediction. In *31st AAAI conference on artificial intelligence*.

Zhang, J. (2017). Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction. *Information Processing & Management*, 53(1), 42-51.

Bojchevski, A., & Günnemann, S. (2017). Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. *arXiv preprint arXiv:1707.03815.*

Wang, Z., Liang, J., & Li, R. (2018). Exploiting user-to-user topic inclusion degree for link prediction in social-information networks. *Expert Systems with Applications*, 108, 143-158.

Wu, L., Fisch, A., Chopra, S., Adams, K., Bordes, A., & Weston, J. (2018, April). Starspace: Embed all the things!. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

Yoon, B., & Magee, C. L. (2018). Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction. *Technological Forecasting and Social Change*, 132, 105-117.

Baek, J. W., & Chung, K. Y. (2021). Multimedia recommendation using Word2Vec-based social relationship mining. *Multimedia Tools and Applications*, 80(26), 34499-34515.

Brochier, R., Guille, A., & Velcin, J. (2019, May). Link prediction with mutual attention for text-attributed networks. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 283-284).

Hassan, D. (2019, July). Supervised Link Prediction in Co-Authorship Networks Based on Research Performance and Similarity of Research Interests and Affiliations. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)* (pp. 1-6). IEEE.

Hettige, B., Li, Y. F., Wang, W., & Buntine, W. (2020, February). Gaussian embedding of large-scale attributed graphs. In *Australasian Database Conference* (pp. 134-146). Springer, Cham.

Ho, T. K. T., Bui, Q. V., & Bui, M. (2019, December). Co-author relationship prediction in bibliographic network: A new approach using geographic factor and latent topic information. In *Proceedings of the Tenth International Symposium on Information and Communication Technology* (pp. 69-77).

Liu, H., Kou, H., Chi, X., & Qi, L. (2019, July). Combining time, keywords and authors information to construct papers correlation graph (S). In *31st International Conference on Software Engineering and Knowledge Engineering* (*SEKE 2019)* (pp. 11-19).

Rahmaida, R., Saefuddin, A., & Sartono, B. (2019). Predicting Potential Co-Authorship Using Random Forest: Case of Scientific Publications in Indonesian Institute of Sciences. *STI Policy and Management Journal*, 4(2).

Tkachenko, M., & Lauw, H. W. (2019, July). Comparelda: A topic model for document comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 7112-7119).

Zhang, Y., Shen, S., & Wu, Z. (2018, August). Improve link prediction accuracy with node attribute similarities. In *International Conference on Computer Engineering and Networks* (pp. 376-384). Springer, Cham.

Zhu, Y., Huang, D., Xu, W., & Zhang, B. (2020). Link prediction combining network structure and topic distribution in large-scale directed network. *Journal of Organizational Computing and Electronic Commerce*, 30(2), 169-185.

Berahmand, K., Nasiri, E., Rostami, M., & Forouzandeh, S. (2021). A modified DeepWalk method for link prediction in attributed social network. *Computing*, 103(10), 2227-2249.

Duricic, T., Kowald, D., Schedl, M., & Lex, E. (2021, November). My friends also prefer diverse music: homophily and link prediction with user preferences for mainstream, novelty, and diversity in music. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 447-454).

Yoon, B., Kim, S., Kim, S., & Seol, H. (2021). Doc2vec-based link prediction approach using SAO structures: Application to patent network. *Scientometrics*, 1-30.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press.

Appendix 1: Topology-based Similarity Measures

Local methods

Common Neighbors (CN):	$CN(u,v) = \lvert\Gamma(u) \cap \Gamma(v)\rvert$
Jaccard Coefficient (JC):	$JC(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\cup\Gamma(v)\rvert}$
Adamic Adar (AA):	$AA(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{log\lvert\Gamma(z)\rvert}$
Resource Allocation (RA):	$RA(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{\lvert\Gamma(z)\rvert}$
Preferential Attachment (PA):	$PA(u,v) = \lvert\Gamma(u)\rvert \cdot \lvert\Gamma(v)\rvert$
Sørensen Index (SO):	$SO(u,v) = \frac{2\cdot\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\rvert+\lvert\Gamma(v)\rvert}$
Salton Index (SI):	$SA(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\sqrt{\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert}}$
Leicht-Holme-Newman Index (LHN):	$LHM(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert}$
Parameter Dependant (PD):	$PD(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{(\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert)^{\lambda}}$
Hub-promoted (HP):	$HP(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{min(\lvert\Gamma(u)\rvert,\lvert\Gamma(v)\rvert)}$
Hub-depressed (HD):	$HD(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{max(\lvert\Gamma(u)\rvert,\lvert\Gamma(v)\rvert)}$
Individual Attraction Index (IA):	$IA(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{\lvert\Gamma(u)\cap\Gamma(v)\cap\Gamma(z)\rvert+2}{\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert}$
Individual Attraction Simplified Index (IA*):	$IA(u,v) = \left(\left\lvert \bigcup_{z\in\Gamma(u)\cap\Gamma(v)} \Gamma(u) \cap \Gamma(v) \cap \Gamma(z)\right\rvert + 2\right) \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert}$
Local Naive Bayes (LNB):	$LNB(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} log\left(\frac{C(z)}{1-C(z)}\frac{1-\rho}{\rho}\right)\ \rho = \frac{2\cdot\lvert E\rvert}{n\cdot(n-1)}$
CAR based Common Neighbor (CCN):	$CCN(u,v) = \lvert\Gamma(u) \cap \Gamma(v)\rvert \cdot \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{\lvert\Gamma(u)\cap\Gamma(v)\cap\Gamma(z)\rvert}{2}$
CAR based Jaccard Coefficient (CJC):	$CJC(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{\lvert\Gamma(u)\cap\Gamma(v)\cap\Gamma(z)\rvert}{2\lvert\Gamma(u)\cup\Gamma(v)\rvert}$

CAR based Adamic Adar (CAA):	$CAA(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{\lvert\Gamma(u)\cap\Gamma(v)\cap\Gamma(z)\rvert}{log\lvert\Gamma(z)\rvert}$
CAR based Resource Allocation (CRA):	$CRA(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{\lvert\Gamma(u)\cap\Gamma(v)\cap\Gamma(z)\rvert}{\lvert\Gamma(z)\rvert}$
CAR based Preferential Attachment (CPA):	$CPA(u,v) = (\lvert\Gamma(u)\rvert - \lvert\Gamma(u)\cap\Gamma(v)\rvert + CCN(u,v)) \cdot (\lvert\Gamma(v)\rvert - \lvert\Gamma(u)\cap$
Local Neighbors Link (LNL):	$LNL(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{\lvert\Gamma(u)\cap\Gamma(z)\rvert+\lvert\Gamma(v)\cap\Gamma(z)\rvert+2}{\lvert\Gamma(z)\rvert}$
Functional Similarity Weight (FSW):	$FSW(u,v) = \frac{2\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)-\Gamma(v)\rvert+2\lvert\Gamma(u)\cap\Gamma(v)\rvert+\lambda}$ $\lambda = max(0, \lvert\Gamma_{avg}\rvert - (\lvert\Gamma(u) - \Gamma(v)\rvert) + (\lvert\Gamma(u)\cap\Gamma(v)\rvert))$
Local Affinity Structure (LAS):	$LAS(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\rvert} + \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(v)\rvert}$
Mutual Information Index (MI):	$MI(u,v) = -I(e_{u,v}\lvert z)$ $I(e_{u,v}\lvert z) = log_2 \frac{\lvert\{e_{u,v}: u,v\in\Gamma(z), e_{u,v}\in E\}\rvert}{\frac{1}{2}\lvert\Gamma(z)\rvert(\lvert\Gamma(z)\rvert-1)}$
Node Clustering Coefficient (CCLP):	$CCLP(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} C(z) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{t(z)}{\lvert\Gamma(z)\rvert(\lvert\Gamma(z)\rvert-1)}$
Node and Link Clustering Coefficient (NLC):	$NLC(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{C(z)}{\lvert\Gamma(z)\rvert-1}(\lvert\Gamma(u)\cap\Gamma(z)\rvert + \lvert\Gamma(v)\cap\Gamma(z)\rvert)$
Weighted Common Neighbors (WCN):	$WCN(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{w(u,z)^{\alpha}+w(v,z)^{\alpha}}{2}$
Weighted Jaccard Coefficient (WJC):	$WJC(u,v) = \frac{\sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{w(u,z)^{\alpha}+w(v,z)^{\alpha}}{2}}{\sum_{u'\in\Gamma(u)} w(u,u') + \sum_{v'\in\Gamma(v)} w(v,v')}$
Weighted Adamic Adar (WAA):	$WAA(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{w(u,z)^{\alpha}+w(v,z)^{\alpha}}{2} \frac{1}{log(1+\sum_{z'\in\Gamma(z)} w(z,z'))}$
Weighted Resource Allocation (WRA):	$WRA(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{w(u,z)^{\alpha}+w(v,z)^{\alpha}}{2} \frac{1}{\sum_{z'\in\Gamma(z)} w(z,z')}$

Evidential Measurement (EM):	$EM(u,v) = \frac{1}{CN(u,v)} \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{1+e^{-a(x+c)}},\ x = \frac{CN(u,v)^2}{PA(u,v)}$
Resource Allocation Common Neighbors Interactions (RACNI):	$RACNI(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\lvert\Gamma(z)\rvert} + \sum_{e}$
Local Interaction Score (LIT):	$LIT(u,v) = \frac{\sum_{z \in \Gamma(u) \cap \Gamma(v)} LIT(z,u)(t-1) + \sum_{z \in \Gamma(u) \cap \Gamma(v)} LIT(z,v)(t-1)}{\sum_{z \in \Gamma(u)} LIT(z,u)(t-1) + \sum_{z \in \Gamma(v)} LIT(z,v)(t-1) + \lambda(u) + \lambda(v)}$ $\lambda(u) = max(0, \sum_{x \in V} \sum_{y \in \Gamma(x)} LIT(x,u)(t)/\lvert V\rvert - \sum_{z \in \Gamma(u) \cap \Gamma(v)} LIT(z,u)(t-1)$

Global methods

Katz:	$Katz(u,v) = \sum_{l=1}^{\infty} \alpha^l \lvert path_l(u,v)\rvert$ In matrix notation, the similarity can be expressed as $Katz(A) = \alpha A + \alpha^2 A^2 + \cdots = (I - \alpha A)^{-1} - I$
Relation strength:	$RS(u,v) = \sum_{l=1}^{L} R_{pl}^{*}(u,v)$ $R_{pl}^{*}(u,v) : \begin{cases} \prod_{k=1}^{K} R(z_k, z_{k+1}),\ K \le r, \\ 0, \quad otherwise. \end{cases}$
FriendLink (FL):	$FL(u,v) = \sum_{l=1}^{L} \frac{1}{l-1} \frac{\lvert A^i(u,v)\rvert}{\prod_{j=2}^{l}(\lvert V\rvert - j)}$
Global Leicht Holme Newman Index:	$GLHN(u,v) = \beta_1 (I - \beta_2 A)^{-1}$
Random Walk (RW):	$RW(u,v) = p(u,v)$
Random Walk with Restart (RWR):	$S_{RWR}(u,v) = q_{uv} + q_{vu}$ where $\vec{q_u} = (1 - \alpha)(1 - \alpha P^T)^{-1} \vec{e_u}$
Pseudoinverse Laplacian Matrix (PLM):	$S_{PLM}(u,v) = L^+_{uv}$ where L^+ is the Pseudoinverse Laplacian Matrix calculated as $L^+ = V\Sigma^+ U^T = \left(L - \frac{ee^T}{n}\right)^{-1} + \frac{ee^T}{n}$

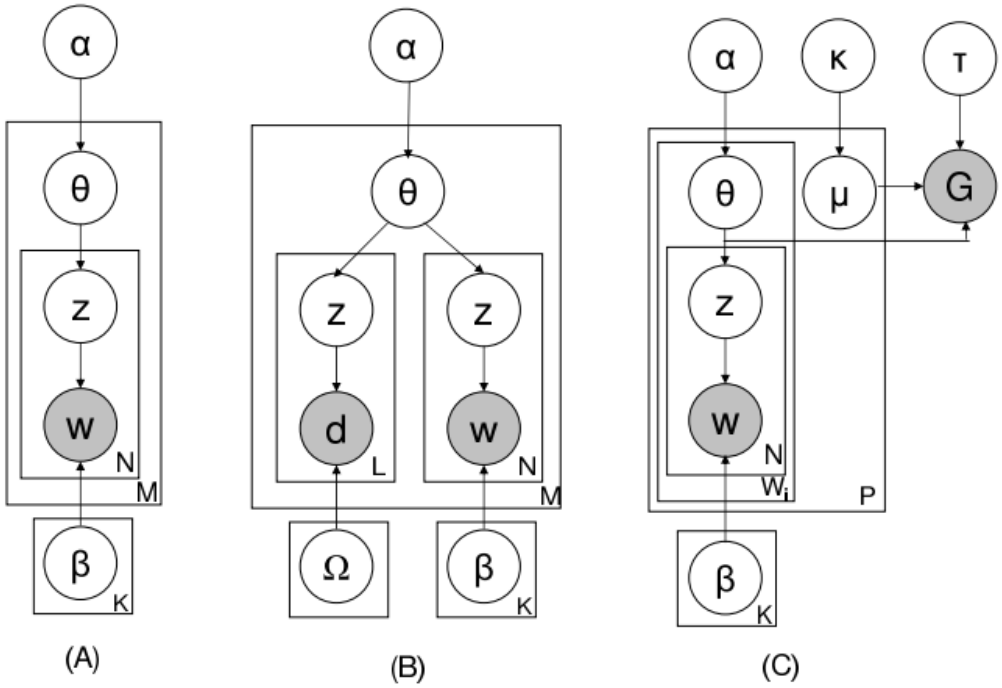
Pseudoinverse Laplacian Matrix Cosine (PLC):	$S_{PLC}(u,v) = \frac{L^+_{uv}}{\sqrt{L^+_{uu} L^+_{vv}}}$ where L^+ is the Pseudoinverse Laplacian Matrix
Hitting Time (HT):	$S_{HT}(u,v) = 1 + \sum_{z \in \Gamma(u)} P_{u,z} S_{HT}(z,y)$ $P = D_A^{-1} A$ where diagonal matrix D_A of A has a value $\left(D_A\right)_{i,i} = \sum_j A_{i,j}$
Commute Time (CT):	$S_{CT}(u,v) = S_{HT}(u,v) + S_{HT}(v,u) =$ $S_{CT}(u,v) = \lvert E \rvert \left(L^+_{uu} + L^+_{vv} - 2L^+_{uv}\right)$ where L^+ is the Pseudoinverse Laplacian Matrix
Average Commute Time (ACT):	$S_{ACT}(u,v) = \frac{1}{L^+_{uu} + L^+_{vv} - 2L^+_{uv}}$ where L^+ is the Pseudoinverse Laplacian Matrix
Normalized Average Commute Time (NACT):	$S_{NACT}(u,v) = \frac{1}{S_{HT}(u,v)\,\pi_v + S_{HT}(v,u)\pi_y}$ where $\pi_u = \frac{\lvert \Gamma(u) \rvert}{\sum_{v \in V} \lvert \Gamma(v) \rvert}$ is the stationary distribution of the Markov Chain describing random walker on the graph.
Matrix Forest Index (MFI):	$S_{MFI} = (I + L)^{-1}$ where $(I + L)_{u,v}$ is the number of spanning rooted forests (u as root) consisting of both the nodes u and v
SimRank (SR):	$S_{SR}(u,v) = \gamma \frac{\sum_{x \in \Gamma(u)} \sum_{y \in \Gamma(v)} S_{SR}(x,y)}{\lvert \Gamma(u) \rvert \lvert \Gamma(v) \rvert}$ where $\gamma \in [0,1]$ is the decay factor
Rooted PageRank (RPR):	$S_{RPR}(u,v) = \left((1 - \beta)(1 - \beta D^{-1} A)^{-1}\right)$
PropFlow Predictor (PFP):	$S_{PFP}(u,v) = S_{PFP}(x,u) \frac{w_{u,v}}{\sum_{z \in \Gamma(u)} w_{u,z}}$
Escape Probability (EP):	$S_{EP}(u,v) = \frac{Q(u,v)}{Q(u,u)Q(v,v) - Q(u,v)Q(v,u)}$ $Q(u,v) = S_{RPR}(u,v)/(1 - \beta) = (1 - \beta D^{-1} A)^{-1}$
Blondel Index (BI):	$S_{BI}(u,v) = S_{u,v}(t = c)$ where $S(t = c)$ refers to the similarity matrix in the steady state level, $S(t) = \frac{AS(t-1)A^T + A^T S(t-1)A}{\left\lVert AS(t-1)A^T + A^T S(t-1)A \right\rVert_F}$

	refers to the similarity matrix at iteration t and$S(0) = I$. $\|\|M\|\|_F$ is the Frobenius matrix norm.
Flow Propagation (FP):	$S_{FP} = D^l A D^r$ where D^l and D^r are diagonal matrices defined as $D^l_{i,i} = 1/\sqrt{\Sigma_j A_{i,j}}$ and $D^r_{i,i} = 1/\sqrt{\Sigma_j A_{j,i}}$ respectively

Quasi Local Methods:

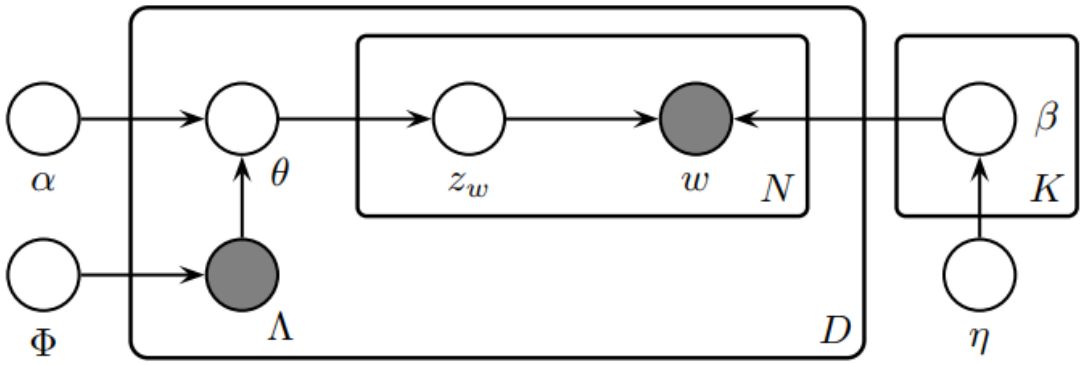
Local path (LP):	$S_{LP} = A^2 + \alpha A^3$
Shortest path (SP):	$S_{SP} = -\|shortest\ path(u,v)\|$
Local Random Walk (LRW):	$S_{LRW}(u,v) = S_{u,v}(t = l)$ Where $S_{u,v}(t = l)$ refers to the similarity matrix after a limited number l of iteration, and $S(t) = \frac{\|\Gamma(u)\|}{2\|E\|} p_{uv}(t) + \frac{\|\Gamma(u)\|}{2\|E\|} p_{uv}(t)$ refers to the similarity matrix at iteration t and $p_{uv}(t)$ is the probability vector obtained by the random walk at iteration t.
Superposed Random Walk (SRW):	$S_{SRW}(u,v) = S_{u,v}(t = l)$ Where $S_{u,v}(t = l)$ refers to the similarity matrix after a limited number l of iteration, and $S(t) = \sum_{i=1}^{t} \left(\frac{\|\Gamma(u)\|}{2\|E\|} p_{uv}(i) + \frac{\|\Gamma(u)\|}{2\|E\|} p_{uv}(i) \right)$ refers to the superposition of the similarity matrix of local random walks across all iterations
Path of Length Three (PL3):	$S_{PL3}(u,v) = \sum_{x,y \in V} \frac{a_{u,x} \cdot a_{x,y} \cdot a_{y,v}}{\|\Gamma(x)\|\|\Gamma(y)\|}$

Appendix 2: Probabilistic Topic Models



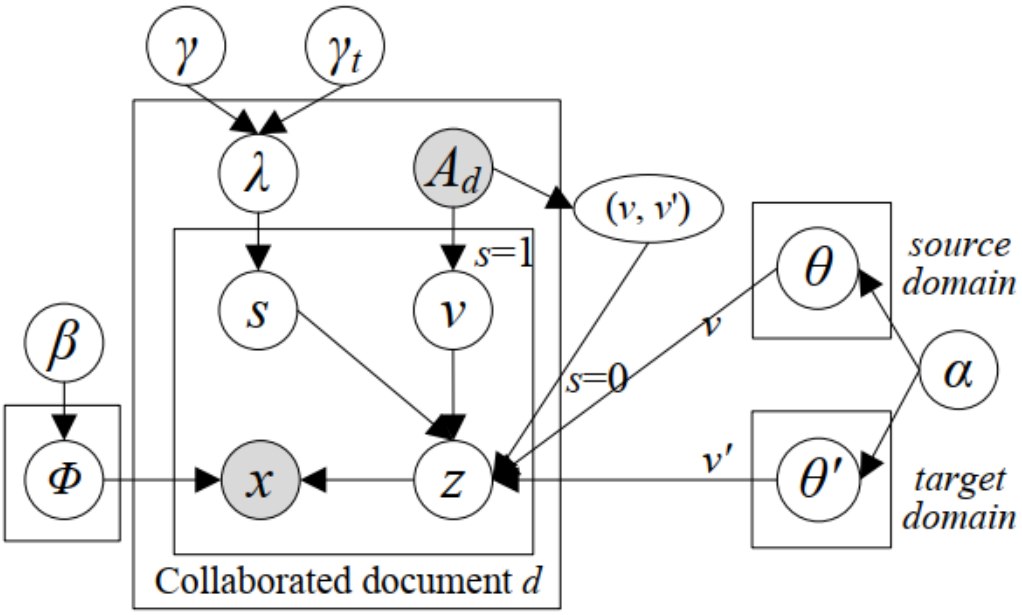
Source: Liu et al., 2009

Figure A2.1: Plate notation diagram of (A) LDA, (B) Link-LDA, and (C) Topic-Link LDA models.



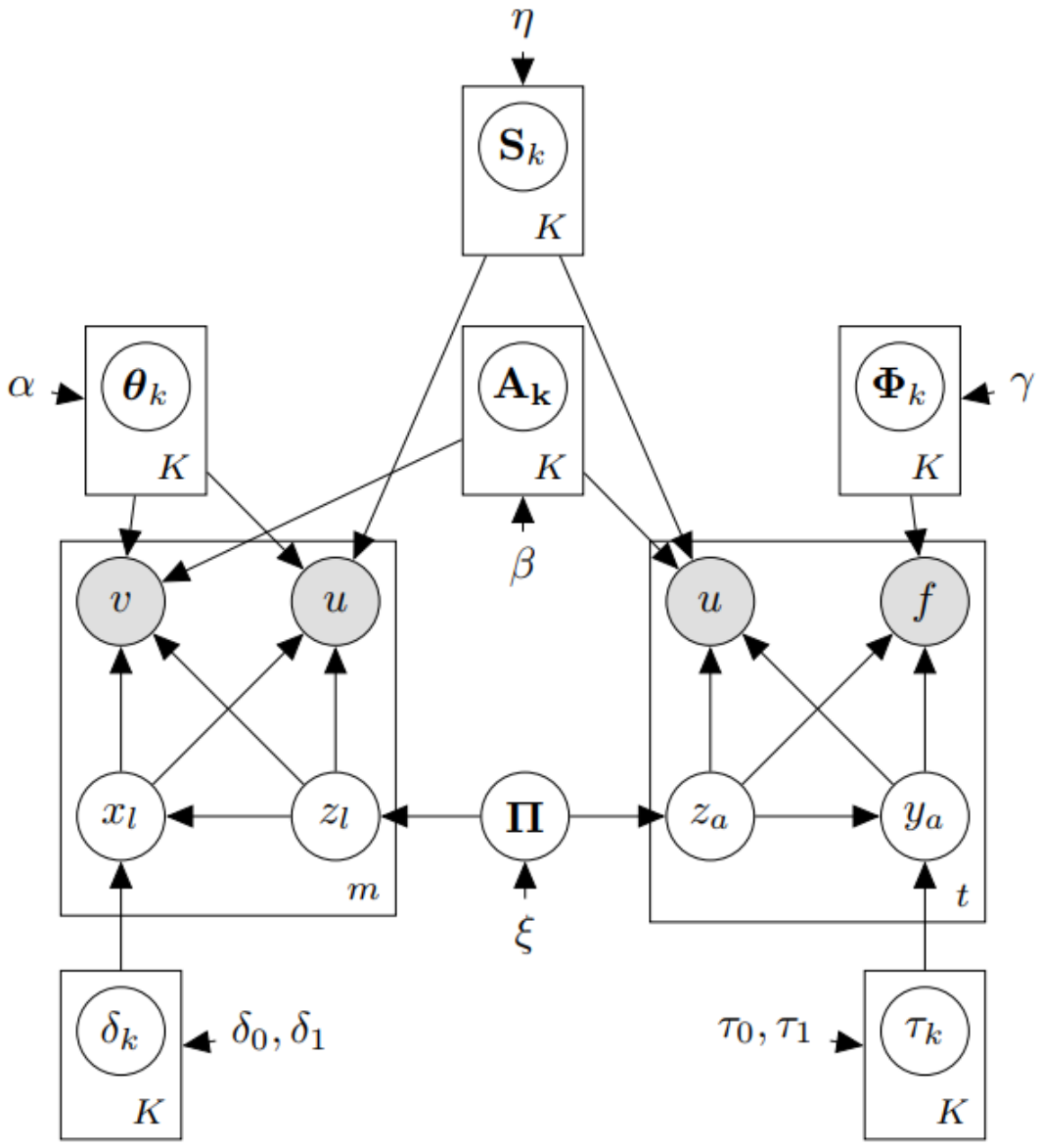
Source: Quercia et al., 2009

Figure A2.2: Plate notation diagram of k Labeled-LDA.



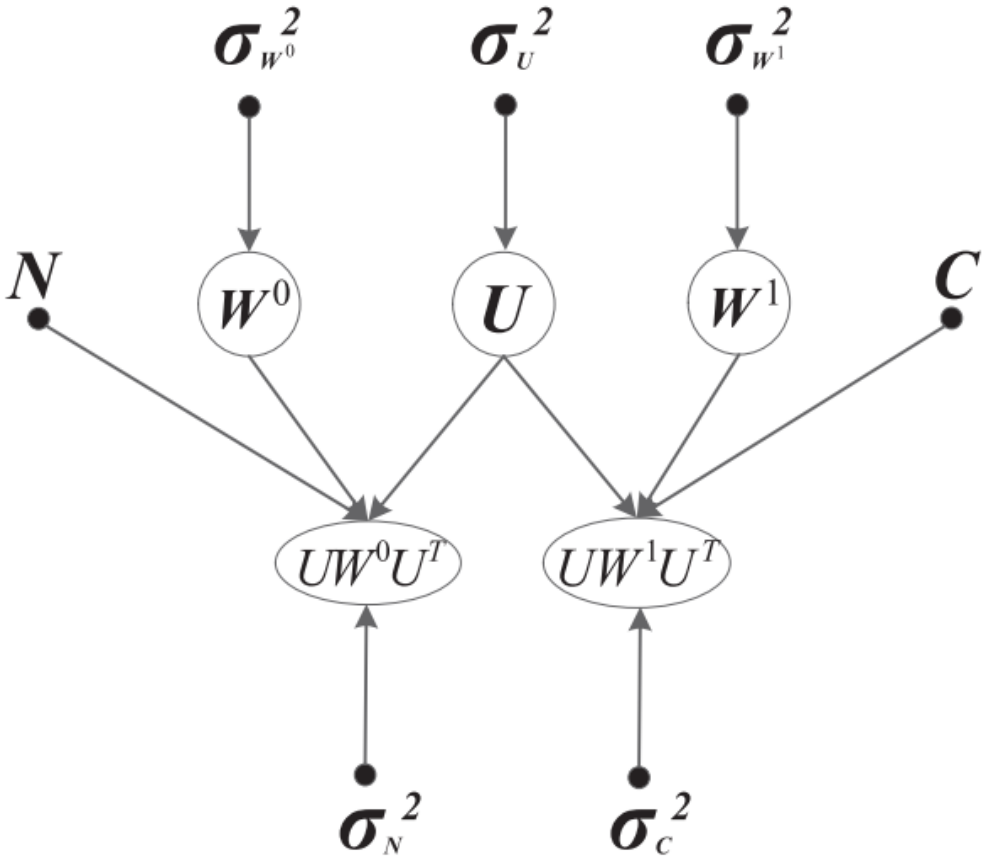
Source: Tang et al., 2012

Figure A2.3: Plate notation diagram of CTL model



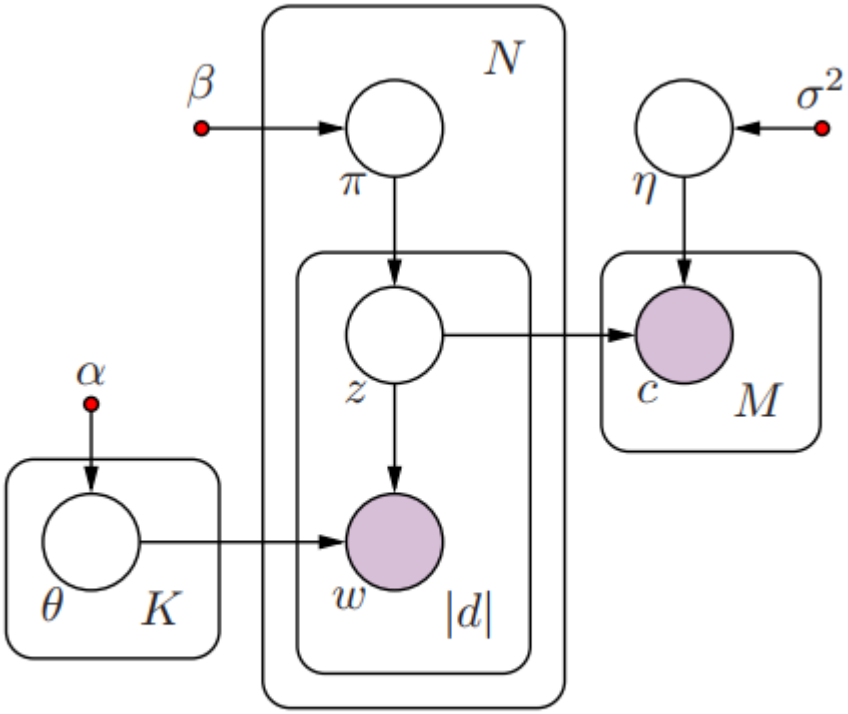
Source: Barbieri et al., 2014

Figure A2.4: Plate notation diagram of WTFW model



Source:
Wang et al., 2018

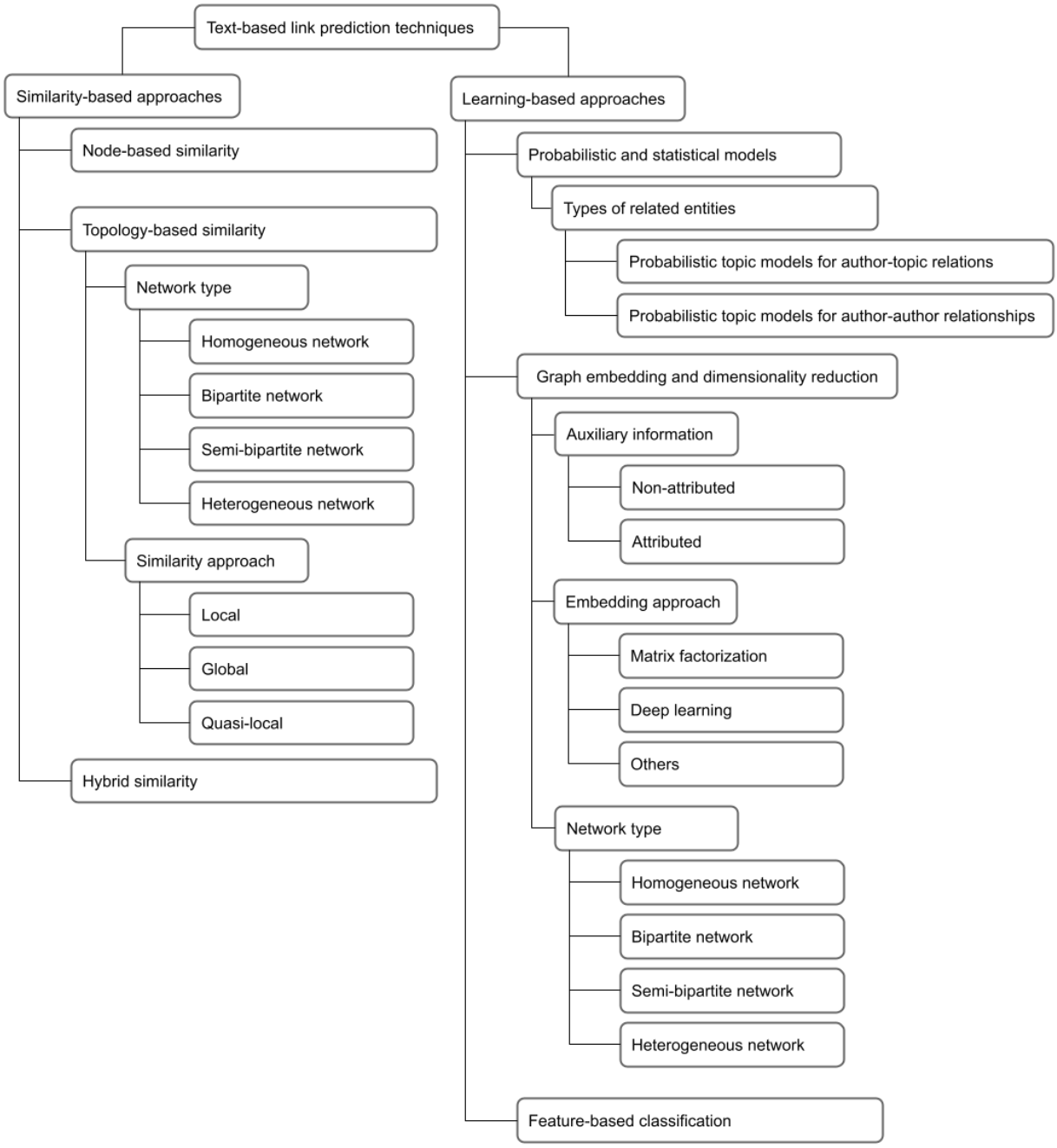
Figure A2.5: Directed probabilistic graphical model representing the relations between the matrices and parameters in FMPF model



Source: Tkachenko-Lauwn., 2019

Figure A2.6: Plate notation diagram of CompareLDA model

Appendix 3: Proposed Taxonomy



CHAPTER II: Augmented link prediction in scientific collaboration networks: Enrichment of node attributes based on correlated topics

1.Introduction

In 2002, Harvard scholars Nicholas Christakis and James Fowler, were introduced to each other for the first time by their mutual friend, political scientist Gary King. This was the beginning of a long and productive collaborative relationship that has made great contributions to our current understanding of social networks. This is an illustrative example of how social networks both enable and restrict our interaction possibilities.

Modeling scientific collaboration as a social network has allowed researchers to make accurate predictions about the formation of new collaborations (Liben-Nowell and Kleinberg, 2007; Hasan and Zaki, 2011; Wang et al., 2015). Inspired by the evidence of homophily in social networks, the main assumption of collaboration prediction is that two scientists who are “close” in the network are more likely to collaborate in the near future (Liben-Nowell and Kleinberg, 2007)

However, from this perspective, collaborative relationships like the one of Christakis and Fowler are particularly difficult to predict using only the information in the network since both scholars were “studying networks from a completely different perspective” as they told in their book, *Connected*, published in 2009. Christakis' interest in networks came from the observation of the influence that health (and illness) of patients have on the people around them, while Fowler was studying the origin of people’s political beliefs and their influence on others.

To deal with this issue, text-based link prediction models have expanded the notion of “closeness” by combining network proximity and interests similarity measures (Hasan et al., 2006; Makrehchi, 2011; Bhattacharyya et al., 2011; Chuan et al., 2017; Rahmaida et al. 2019, Zhang et al., 2019). In this paper, we present a simple method that leverage state-of-the-art text mining models to augment text-based link prediction techniques. Instead of focusing on interests matching, our approach looks into how scholars’ interest on correlated topics can help us capture new types of collaboration.

Two decades after Christakis and Fowler met, advances in the field of network science have allowed to uncovered complex structural patterns and dynamics of self-organization in scientific collaboration networks (Newman, 2001; Barabasi et al., 2002; Wagner and Leydesdorff, 2005; Molontay and Nagy, 2020). Identifying the fundamental laws of the evolution of the scientific collaboration network is key to understanding the underpinning mechanisms of team formation in science (Zeng et al., 2017).

The attention on this matter has been fueled by the increasing dominance of teams in scientific production becoming a central question at the heart of the study of science from the perspective of complex systems (Fortunato et al., 2018). This trend is reflected not only in the higher proportion of papers published by teams (Abt, 2007) but in the higher impact (Wuchty et al., 2007) and quality of their research (Franceschet and Costantini, 2010). Research impact was measured as the number of citations received (Wuchty et al., 2007), while research quality was judged qualitatively by peer reviewers (Franceschet and Costantini, 2010).

Predicting future collaborations addresses the challenge of assessing theoretical models of team formation using real word data (Liben-Nowell and Kleinberg, 2007). Traditionally, mainstream link prediction research aims to develop and improve general purpose techniques relying exclusively on features derived from the network topology. This approach is capable of achieving decent performance regardless of the context in which it is applied. As a consequence, link prediction techniques in social networks have evolved rapidly into a wide variety of methods and applications (Pandey et al. 2019; Samad et al. 2020; Yuliansyah et al., 2020).

However, several studies have reported that combining topology-based methods with text-based attributes can significantly improve the link prediction performance in scientific collaboration networks (Hasan et al., 2006, Wang et al., 2007, Wohlfarth and Ichise, 2008; Sachan and Ichise, 2011). In this direction, recent developments are becoming increasingly sophisticated and complex by combining text-based attributes with machine learning models and deep neural network approaches (Pham et al., 2021; Shu et al., 2021; Wang et al., 2021; Xu et al., 2021)

In contrast, we argue that there is still a gap in how text-based attributes are constructed. Better designed attributes could clearly be a complement and improve the performance of more sophisticated models. With the example of Christakis and Fowler in mind, our work aims to fill this gap by including information about correlated topics into the text-based attributes to enrich the representation of research interests. The contributions of our work are threefold. First, we elaborate on how text-based attributes and topic relatedness have been applied in the field of link prediction in scientific collaboration networks. Second we propose a simple method to incorporate information about correlation between topics into the representation of nodes attributes. And third, we provide an empirical example of the proposed approach in comparison to a set of baseline methods.

The outline of this paper is as follows. In Literature Review, we conduct a brief review of previous research in the area of text-based methods for link prediction in scientific collaboration networks. In the Proposed Approach, we provide the details of the proposed model. In Experimental Procedure, we describe the source of data, the experimental setup, and evaluation procedures. In Results and Discussions, we present results of the evaluation of the model in comparison to baseline methods and reflect on the findings. In Conclusions, we summarize our findings and discuss their implications.

2.Literature Review

2.1.Text-based attributes for link prediction

Previous works have shown that metadata associated with networks can be incorporated into link prediction methods as node attributes, significantly improving the prediction performance (Hasan et al., 2006). In scientific collaboration networks, bibliographic databases are great sources of document metadata (Zeng et al., 2017).

In most cases, scientific publications' metadata like the article title, keywords and abstract are informative about the research content of the paper. Therefore, the research interest of scholars can be modeled by the content of their publications (Chaiwanarom and Lursinsap, 2015). Different text-processing techniques have been used to transform text and labels into an author-attribute matrix.

Hasan et al. (2006) represented authors attributes as bag-of-words, this is lists with all the words present in the authors papers. This representation is equivalent to a author-attribute matrix X where attributes are individual keywords. This approach has shown positive effect on performance when combined with topological measures to train machine learning models like Support Vector Machine, Decision Tree, K-Nearest Neighbors (Hassan et al., 2006), (Wolfarth and Ichise, 2008), and Random Forest (Rahmaida et al., 2019).

Another widespread text-representation method is the term frequency-inverse document frequency matrix (TFIDF). This matrix represents the relevance of a word in a document as a combination between its frequency in the document and its rarity in the corpus (Gentzkow, Kelly and Taddy, 2019). This methods has been used to model the authors' attributes X and to compute similarity between two authors using the cosine similarity (Wang et al., 2007; Bartal et al., 2009; Zhang et al., 2014), and the L1-distance (Sachan and Ichise, 2010). Other applications of TFIDF in link prediction includes pre-processing the document-term matrix to feed learning based algorithms like attributed graph embedding (Hettige et al., 2020) and Doc2vec (Yoon et al. 2021).

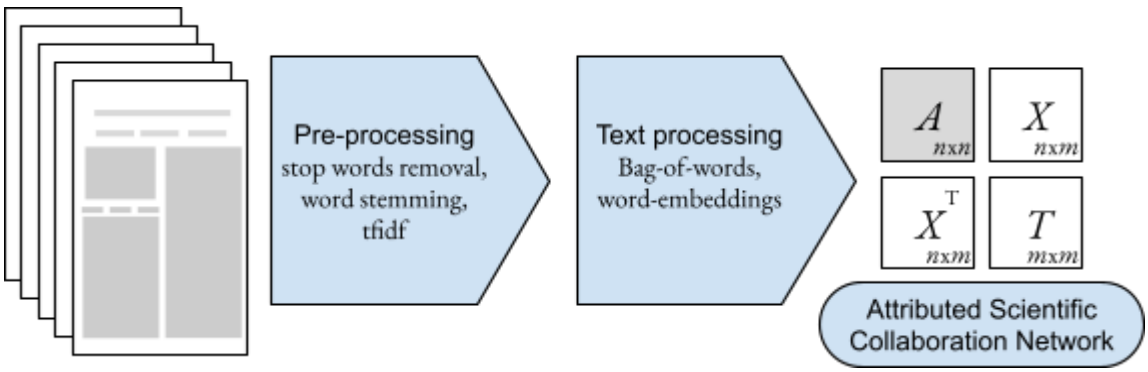


Figure 1: Diagram of attributed scientific collaboration networks

Probabilistic topic models represent documents as probability distributions over a set of latent topics. The most prominent method of this kind is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). In link prediction, the probability distributions obtained from LDA and other topic models have proved to be useful to model the author-attribute matrix X. These attributes are in turn used to compute similarity between pairs of authors using measures like the Hellinger distance (Ho, Bui and Bui, 2019) and cosine similarity. LDA-based cosine similarity has also been used in hybrid similarity measures by combining content similarity with topology-based similarity measures. (Solaimannhezad, 2017; Chuan et al., 2017).

All these methods share the implicit assumption that node attributes are independent of each other. Distance-based similarity measures adopt a space vector framework were authors' attributes are represented by K-dimensional vectors that in practice assume independence of the orthogonal basis and do not account for term correlations (Wong et al., 1987). The use of the Dirichlet distribution in LDA also entails assumptions about the independence of the latent topics.

2.2. Topic correlation in text-based link prediction

Positive correlations between two topics indicate that both topics are likely to co-occur in some documents. For example, scientific articles in the field of behavioral economics are likely to discuss topics about bounded rationality and decision making, while works in neuroeconomics may do the same with topics about decision making and brain activity. These correlations carry relevant information for modeling how interactions between scholars may be modulated not only by their research topics but also by how these topics are related to each other.

One way to represent the topic interactions is the topic-topic adjacency matrix T. In Hassan et al. (2006) such matrix is built based on the co-occurrence of keywords in any paper. Then, they combine the author-author matrix A, the author-keyword matrix X, and the keyword-keyword matrix T into a block matrix representing the entire attributed scientific collaboration network and calculate the similarity between two authors using topology-based measures.

A similar approach is used by Makrehchi (2011) where the author-attribute matrix X is calculated directly as the projection of the LDA-based author-attribute matrix X, into a topic co-occurrence matrix $T = X^T X$. Then, the similarity between authors is computed using topology-based measures on a semi-bipartite network where the author-author relations are unknown.

The incorporation of topic correlations in link prediction have also been considered in other contexts like online social networks. In Bhattacharyya et al. (2011) the topic-topic matrix T is generated based on the semantic relationships of words from a large lexical database named WordNet (Fellbaum 1998). The forest generation heuristics consider relations like hypernyms and hyponyms, holonyms and meronyms, and synonyms and similars. The result it's a network composed of multiple hierarchical trees of related words. Similarity between authors within this

framework is calculated as a hybrid measure that combines distance in the topic-topic network and topology-based measures in the author-author network.

In Zarrinkalam et al. (2016) topics are derived from tweets' content and the relation topics is determined by a combination of Wikipedia-based relatedness measure (Witten et al., 2008) and an algorithm based on collaborative filtering.

3.Proposed Approach

The incorporation of topic correlation in link prediction methods has been addressed in the context of scientific collaboration networks and online social networks under the same principles: Expand the authorship network to include author-topic and topic-topic relations, to then apply well known topology-based methods.

Although this approach has reported improvements in performance to the best of our knowledge no previous work has attempted to extract information from the topic correlations to enrich the authors attributes. The potential benefits of this approach are the simplification of similarity calculation, and flexibility to combine the attributes with other, more sophisticated learning-based link prediction methods like graph embedding among others.

We propose a new approach on the basis of transforming the outcome of a topic model named structural topic model (STM) (Roberts et al., 2013), to enrich the author-topic data with information encoded within the relationship between correlated topics. Our approach consist of three steps depicted in Figure 2 and explained in detail as follows.:

1. Estimate the structural topic model and compute the author-topic matrix X_{AT}.
2. Compute the correlation matrix between the latent topics of the STM model and transform it into binary values.
3. Summarize the author attributes and the topic correlation data into the new author-attribute matrix.

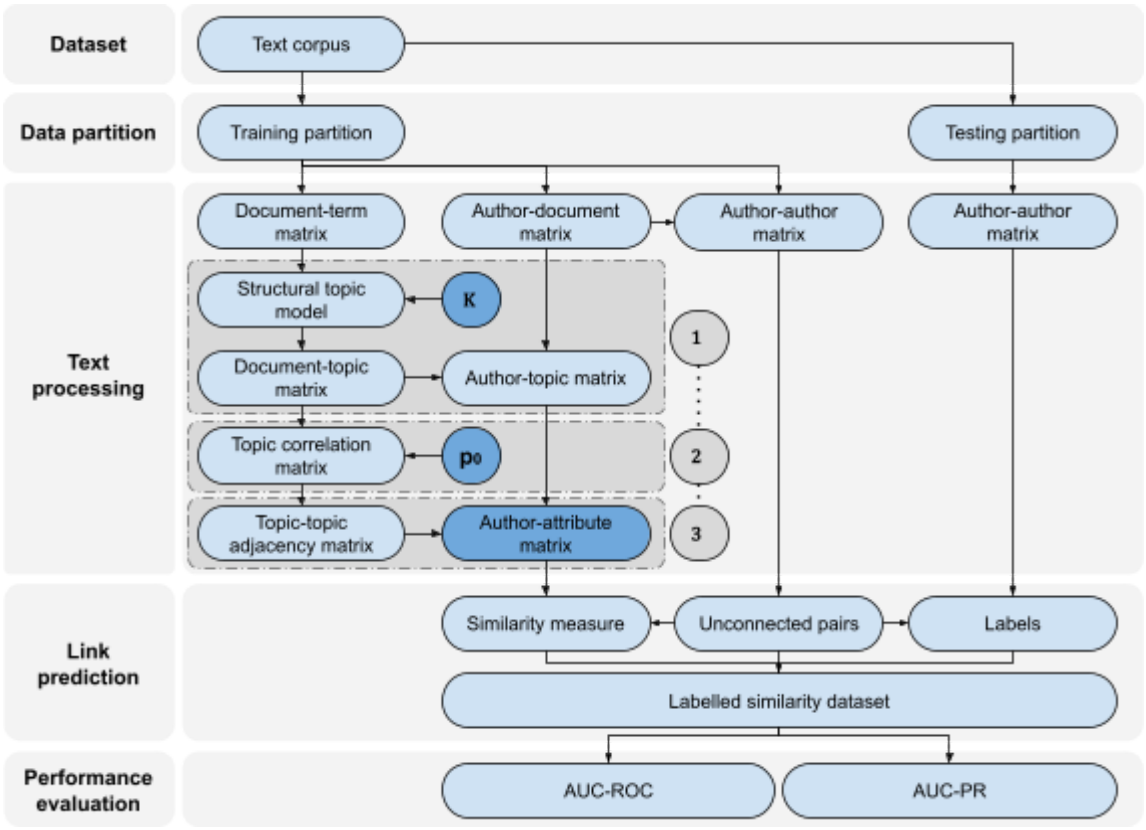


Figure 2: Diagram of the proposed approach

3.1. Structural topic model and author-topic matrix

Probabilistic topic models, such as latent Dirchlet allocation (LDA) (Blei et al., 2003) and Correlated Topic Model (CTM) (Blei and Lafferty, 2007) have been extensively used to model an explore scientific literature (Chen et al., 2017).

Structural topic model (STM) is a flexible topic model that permits users to incorporate documents' metadata into the topic models as covariates (Roberts et al., 2013). Without the inclusion of covariates, STM is equivalent to a faster and more accurate implementation of logistic-normal topic model than the CTM model (Roberts et al., 2019).

The advantage of using the logistic-normal distribution over the traditional Dirchlet distribution used by LDA is that it is not limited by independence assumptions allowing correlations between the latent topics. This gives a more realistic model of latent topic structure, and in the case of STM to incorporate and estimate the effect of covariates on the topics' prevalence and content.

Similarly to LDA, the outcome of the STM model comprises two probability matrices: One containing topic distributions on words, denoted by X_{TW}, and the other containing the document distributions over topics, denoted by X_{DT}.

By taking the document-topic matrix X_{DT} and the author-document matrix X_{AD}, that represents the authorship relations between authors and documents, the author-topic matrix can be computed as:

$$X_{AT} = diag(A)^{-1} \cdot X_{AD} \cdot X_{DT}$$

where A is the author-author adjacency matrix obtained as $A = X_{AD} \cdot X_{AD}^{T}$, consequently the values of the diagonal $a_u = A_{u,u}$ represent the total number of documents published by the author u. It follows that values in each row x_u of the author-topic matrix X_{AT} corresponds to the average, for each topic, of the topic distributions of the documents published by the author u.

3.2. Topic-topic adjacency matrix

The document-topic matrix obtained from the STM model is used to calculate the topic correlation matrix. In order to transform the topic correlation matrix into an adjacency matrix with binary values we use a correlation threshold ρ_0. If two topics are correlated above that threshold, then those two topics are considered to be connected in the topic-topic network (Roberts et al., 2014).

3.3.Summarizing correlated topics

The attribute summarizing function $f: [0,1]^2 \rightarrow [0,1]$ defines how the authors' attributes are combined according to the relations in the topic-topic adjacency matrix T, which is defined as positive with values ranging from 0 to 1. We explore two alternative domains for the function, and three options for the summarizing operation (see Figure 3).

Domain 1: f is defined only for the lower triangle of T, excluding the matrix diagonal.

$$f(x_u) = f_s(x_{u,i}, x_{u,j}), \;\; if\; T_{i,j} = 1,\; i > j$$

Domain 2: f is defined for both the lower triangle and diagonal of T matrix

$$f(x_u) = \begin{cases} x_{u,i} & , if\; T_{i,j} = 1,\; i = j \\ f_s(x_{u,i}, x_{u,j}) & , if\; T_{i,j} = 1,\; i > j \end{cases}$$

While domain 1 only generates new attributes, domain 2 combines the new attributes with the author-topic matrix X_{AT}, preserving the original values obtained from the structural topic model.

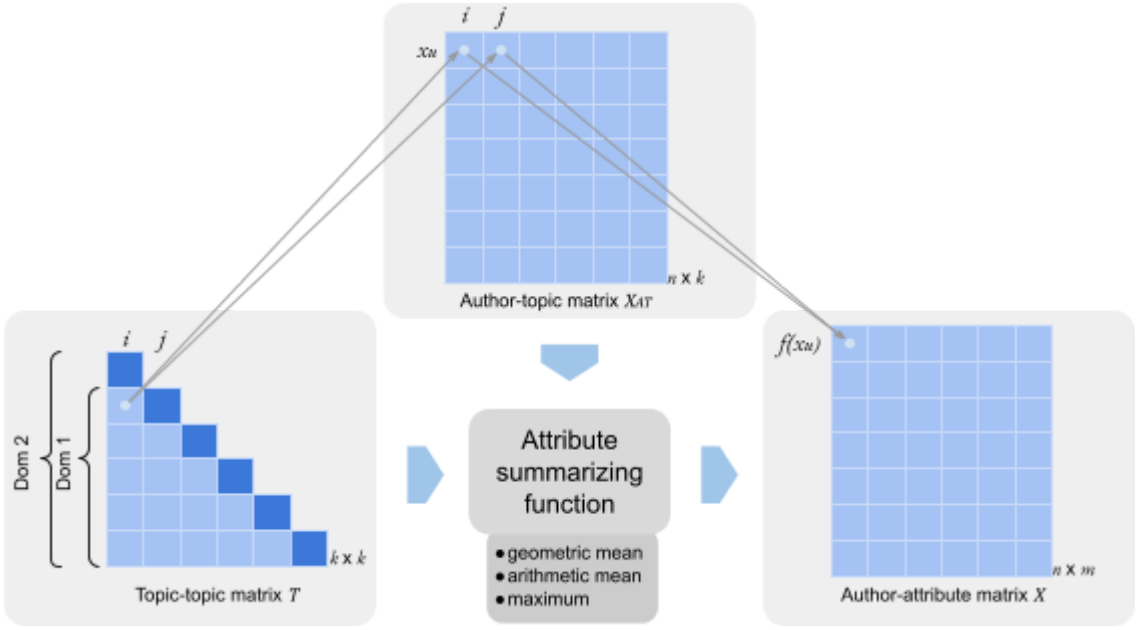


Figure 3: Diagram of the attribute summarizing function

For each pair of correlated topics $< i, j >$, the summarizing operator f_s receives the corresponding values $x_{u,i}$ and $x_{u,j}$ of the topic-based attribute vector x_u and return a value between 0 and 1 that represents the author's interest in the semantic subfield represented by that specific topic combination $< i, j >$. We explore three summarizing operations, namely: Geometric mean, arithmetic mean and maximum. Each one of the operations have a different interpretation as explained in Table 1. $x_{u,<i,j>}$

Table 1: Summarizing operations

Operation	Formula	Interpretation
Geometric mean	$x_{u,<i,j>} = f(x_{u,i}, x_{u,j}) = \sqrt{x_{u,i} \cdot x_{u,j}}$	**Selective**: If the author u is not interested at all in one of the topics neither would be in the joint topic $< i, j >$.
Arithmetic mean	$f_2(x_{u,i}, x_{u,j}) = (x_{u,i} + x_{u,j})/2$	**Moderate**: The interest of the author u in the joint topic $< i, j >$ it is not determined by a single topic but the both contribute equally.
Maximum	$f_3(x_{u,i}, x_{u,j}) = max(x_{u,i}, x_{u,j})$	**Enthusiast**: It is enough that author u to be interested in one of the two topics for to be also interested in the joint topic $< i, j >$ with the same strength.

4.Experimental Procedure

4.1.Dataset

In order to test the robustness of the proposed method in different scenarios we selected collaboration networks from three scientific fields, namely: economics, psychology, and neuroscience. The dataset was retrieved from the OpenAlex bibliographic catalog API. The queries were generated to include all documents with the search term that were published in journals within a 10-year period (2012-2021). The search resulted in three corpora containing around 165.000 documents in the aforementioned fields.

OpenAlex is a fully-open source of scholarly metadata, containing 100% open data, open API, and open-source code (Priem, Piwowar, and Orr, 2022). The OpenAlex data model considers five types of entities: works, authors, venues, institutions, and concepts. Regarding disambiguation, OpenAlex ID is built on the basis of Canonical External ID (CEID) for example, using DOI (digital object identifier) for works, and ORCID for authors. Works' metadata provides rich text data like the document title, concepts, keywords and the abstract inverted index.

Table 2: Description of scientific collaboration networks

	Economics	Psychology	Neuroscience
Number of papers	73436	67397	24416
Number of authors	135860	124502	63880
Mean papers per authors	1.20	1.22	1.38
Mean authors per paper	2.23	2.25	3.6
Mean collaborators per author	3.18	3.23	6.84
Network density	7.2×10^{-5}	8.0×10^{-5}	36.2×10^{-5}
Size of giant component	6421 (5%)	17420 (26%)	23960 (38%)
Size 2nd largest component	3108	207	35
Mean distance	13.5	11.4	7.21
Maximum distance	40	36	26
Clustering coefficient C	0.91	0.89	0.74

The structure of the collaboration network of each selected field is summarized by using standard descriptive analysis (Newman, 2001). The results in Table 2 show some difference between the three networks that are important to point out.

The field of economics presents a small-size ratio between the entire network and the giant component (i.e. the largest connected component in the network) indicating a sparse structure composed of a large number of unconnected communities. This is reflected by a two-mode distance distribution with local maximum in 6 and 17, as shown in Figure 4. Additionally, within the first component the high clustering coefficient and high distances are signs of dense communities but poorly connected between them.

In spite of having similar numbers of papers, authors, papers-per-author, authors-per-paper and collaborators-per-author, psychology and economics collaboration networks largely differ in their structure. Psychology network structure has a considerable giant component equivalent to approximately 26% of the network size.

Research works in neuroscience are conducted by larger teams leading to lower network distances (i.e. the length of the shortest path between two nodes), and a larger giant component close to the 38% relative to the size of the whole network. The number of collaborators per author doubles the amount of economics and psychology, this also increases the number of possible triangles affecting the clustering coefficient.

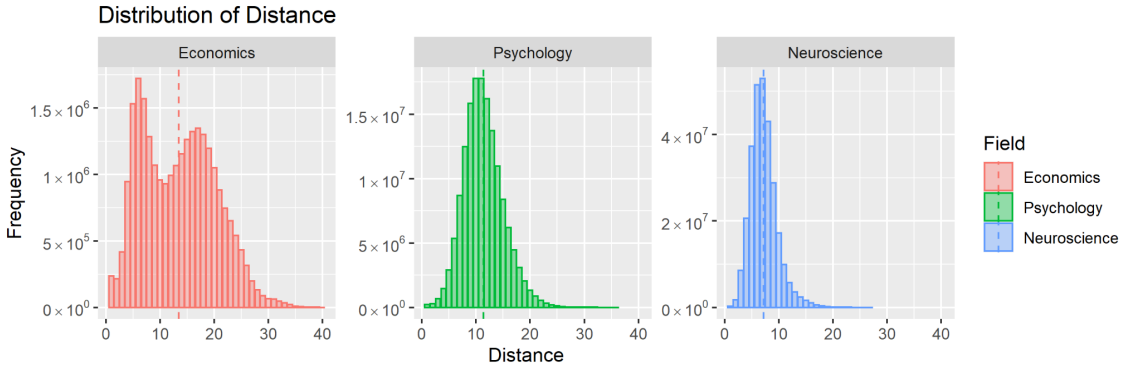


Figure 4: Distribution of network distance

4.2. Data partition

In order to calculate the proposed similarity measures we adopt an unsupervised link prediction approach. For this, we divide the data based on document publication date into a training partition of the network $G_0 = \langle V_0, E_0 \rangle$ corresponding to the period $T_0 = [2012, 2019]$, and a testing partition $G_1 = \langle V_1, E_1 \rangle$ corresponding to the period $T_1 = [2020, 2021]$. Only the training

documents are taken into account for the calculations of the node attributes and the similarity between nodes.

Table 3: Size of training and testing partitions

	Economics		Psychology		Neuroscience	
Partition	Training	Testing	Training	Testing	Training	Testing
Documents	58622 (80%)	14814 (20%)	52803 (78%)	14595 (22%)	19824 (81%)	4592 (19%)
Nodes	104454 (74%)	35787 (26%)	94153 (73%)	34939 (27%)	50316 (75%)	16418 (25%)
Links	494057 (74%)	170614 (26%)	451243 (73%)	169720 (27%)	554273 (75%)	184524 (25%)

Table 3 shows the resulting number of documents, nodes, and links in the training and testing partition for each field. The partition ratio between the number of documents is around 80% : 20% while the number of nodes and links is close to 75% : 25%.

4.3. Text-processing

The proposed approach is compared to a baseline of some of the most popular text processing methods available to generate text-based attributes. In Table 4, we list the baseline methods used in the comparison.

Let K be the number of attributes in the author-topic attribute matrix X_{AT} (i.e. the number of columns). For each method, we calculate X_{AT} with K taking two values: 100 and 200. In the case of topic models LDA and STM, the parameter K represents the number of topics, and in NMF and word-embedding methods, represents the model dimensions. In the case of TFIDF, we defined the parameter K to indicate the top-K words arranged by their TFIDF value.

Table 4: Text-processing methods selected for baseline

Text representation	Text-processing methods		Reference
Bag-of-words	TFIDF	Term frequency - Inverse document frequency	(Luhn, 1957; Spärck, 1972)
	NMF	Non-negative matrix factorization	(Lee and Seung, 1999)
	LDA	Latent Dirichlet Allocation	(Blei, Ng, and Jordan, 2003)
	STM	Structural Topic Model	(Roberts, Stewart, and

			Tingley, 2013)
Word-embeddings	Word2Vec	Continuous bag of words	(Mikolov et al., 2013)
	Glove	Gloval Vectors	(Pennington, Socher, and Manning, 2014)

4.4.Link prediction approach

For the implementation of the link prediction task, we adopt a similarity-based approach. Particularly, for each pair of scholars $< u, v >$ we calculate the cosine similarity between their respective text-based attributes vectors x_u and x_v as follows:

$$S_{COS}(u,v) = \frac{x_u \cdot x_v}{||x_u|| \cdot ||x_v||} = \frac{\sum_{i=1}^{K} x_{u,i} \cdot x_{v,i}}{\sqrt{\sum_{i=1}^{K} x_{u,i}^2 \cdot \sum_{i=1}^{K} x_{v,i}^2}}$$

This measure represents the cosine of the angle between vectors x_u and x_v. And it has been applied extensively in the fields like scientometrics and library science to measure the similarity between documents.

In the particular context of text-based link prediction methods, the cosine similarity measure has been used with attributes based on different text-processing approaches like bag-of-words (Hasan, 2006; Liu et al., 2019), TFIDF matrix (Bartal et al., 2009; Zhang and Yu, 2014; Wang et al. 2017), NMF (Duricic et al., 2021), Word2Vec (Baek and Chung, 2019), and LDA (Pham and Du, 2021).

In most studies the cosine similarity improved the overall classification performance and ranked among the most critical features in the link prediction task (Wang et al., 2007; Bartal et al., 2009; and Zhang and Yu, 2014). Aiello et al. (2012) compared different similarity measures for link prediction tasks in online social networks data (i.e. aNobii, and Last.fm) and observed that cosine similarity performed the best. Zhang and Yu, (2014) observed that cosine similarity between $tfidf$ attributes was the second best rated among 12 topological and non-topological features.

4.5.Performance evaluation

The link prediction evaluation consists of comparing the prediction generated from the available data in the training partition T_0, to the ground truth given by the state of the network during the testing period T_1. For practical reason, instead of making predictions for each possible pair of nodes not yet present in G_0 (i.e. $|V_0| \times |V_0| - E_0$) we construct a labeled dataset consisting of

positively and negatively labeled pairs of authors. The labels indicates a link will or not between the pair of nodes a link during the testing period T_2.

For the positive class we included all the pairs that are not connected in G_0 and connected in G_1, and for the negative class we randomly choose pairs of authors that are not connected in G_0 nor in G_1. With the intention to preserve to some degree the natural class imbalance we impose a sample size for the negative class 10 times the number of pairs in the positive class, as shown in Table 5.

Table 5: Size of positive and negative classes

Sample size	Economics	Psychology	Neuroscience
Positive class (n_{y+})	1097	2414	4372
Negative class (n_{y-})	10970	24140	43720

For each pair of nodes in the labeled dataset we calculate the similarity measure and use it to calculate two curve-based performance measures, namely: the area under the curve (AUC) in the Receiver-Operator characteristic curve (ROC) and in the Precision-Recall curve (PRC). The ROC curve graphically represents the trade-off between the detection of positive cases (TPR) and the false positive rate (FPR). The value of the area under this curve ranges from 0 to 1. While the PR curve represents the trade-off between prediction and recall. The area under the PR curve is particularly useful and informative when applied to binary classification tasks in imbalanced datasets. A higher value also indicates a better classification. In both measures, a higher value indicates a better classification

4.6.Hardware and Software

All statistical analysis was implemented using R statistical software and RStudio integrated development environment (IDE). The hardware features of the experimental environment are described in Table 6.

Table 6: Hardware and software features of the experimental setup

Item	Detail
OS	Microsoft Windows 10 Home
CPU	Intel(R) Core(TM) i7-7700HQ
RAM	32Gb
GPU	Intel(R) HD Graphics 630

5.Results and Discussion

This section presents empirical results of the proposed approach. First we analyze the model performance depending on the parameters values and design decisions. Then, we evaluate the model performance in comparison to the baseline methods and discuss the findings implications.

5.1.Model optimization and parameter tuning

As described in a previous section the outcome of the proposed approach depends on the number of topics K, the correlation threshold ρ_0, the selected domain and summarizing function.

The following results show the effect of these factors in the overall performance of the model in the link prediction task.

5.1.1.Selection of the number of topics (K)

In general, the number of topics in probabilistic topic models is a fixed number specified by the user and does not follow a rule of thumb. Existing data-driven tools can guide the decision making for the number topics by performing automated tests (Roberts et al., 2019). These tests include held out likelihood (Wallach et al., 2009), residuals analysis (Taddy, 2012), semantic coherence (Mimno et al., 2011), and lower bound analysis (Tzikas, 2008).

We selected random samples of 20% and 50% of the corpus to test models ranging from 10 to 300 topics and for the whole corpus from 50 to 300 topics. The test results provide enough evidence to narrow the selection range to values of K between 100 and 200 topics.

Held-out likelihood uses the document completion method to hold out some fraction of the words in a set of documents, train the model and use the document-level latent variables to evaluate the probability of the held-out portion (Wallach et al., 2009). The results in both samples show higher held-out likelihood as the number of topics increases but for the whole corpus this effect is not distinguishable. Only a small hump can be seen in economics and psychology corpora within the range between 100 and 250 topics. In the case of neuroscience variations are barely appreciable.

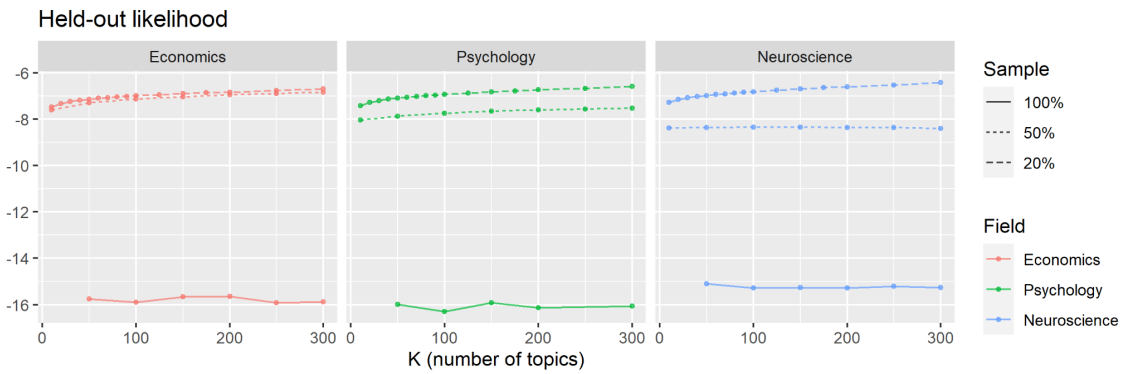


Figure 5: Held-out likelihood by number of topics and sample size

Residuals diagnostic returns the estimated sample dispersion of the residuals. When the model is correctly specified under the data generating process this measure equals the theoretical multinomial dispersion: $\sigma^2 = 1$. Values greater than one imply that the latent topics cannot account for the overdispersion and the number of topics should be increased (Taddy, 2012). The results of the model diagnostic for both samples (20% and 50%) in economics and psychology corpora show U-shaped curves with minimum values between 50 and 200 topics. In the case of neuroscience both curves reach their minimum at 300 topics. Regarding the models fed with 100% of the documents, only the field of economics showed a U-shaped curve with minimum values between 100 and 200 topics. Psychology and neuroscience curves did not show a clear pattern.

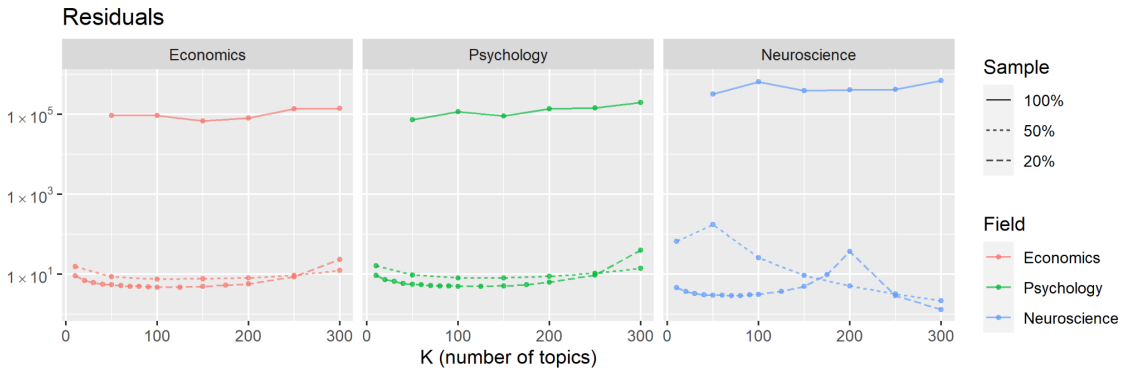


Figure 6: Residuals dispersion by number of topics and sample size

Semantic coherence is the frequency with which the most probable words in a given topic co-occur together in the same documents. This metric is a good predictor of topic quality scores given by expert evaluation (Mimno, 2011). Since a low number of topics can easily reach high levels of semantic coherence, it is recommended to use this metric in a trade-off with other quality indicators (Roberts et al., 2014). The results show that curves' elbow consistently occur

around 100 topics in all three fields for both sample sizes, but when using the whole corpus the semantic coherence of the models is not affected by the number of topics.

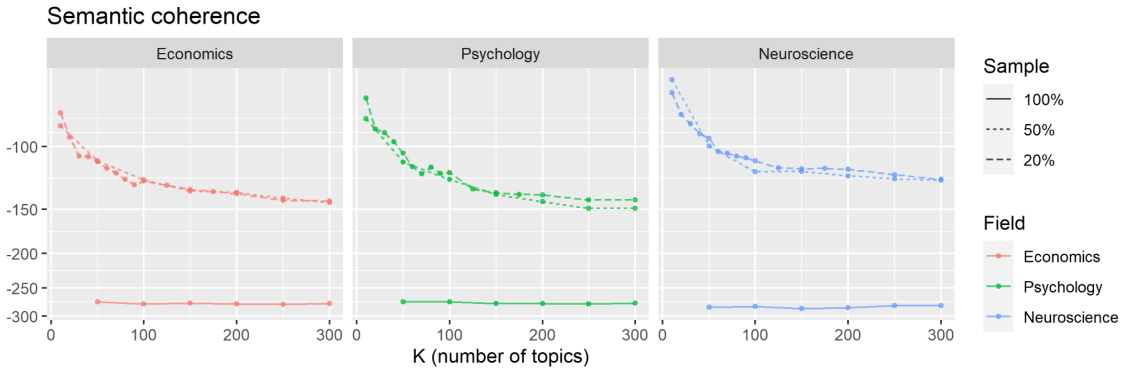


Figure 7: Semantic coherence by number of topics and sample size

Lower bound it is part of the expectation-maximization algorithm and in variational inference. The maximization of the lower bound is directly related to a higher probability of observing all data. Estimated models with higher bounds are preferred (Roberts et al., 2014). The results show consistent curves with higher lower bound as the number of topics increases.

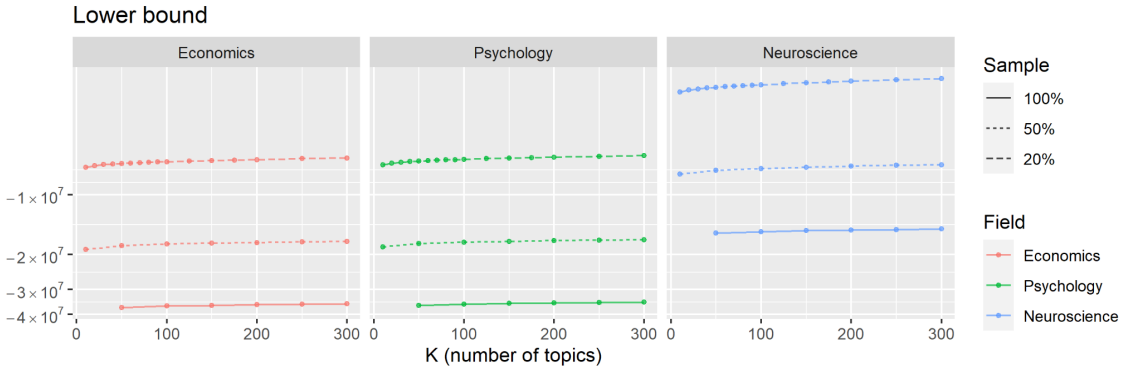


Figure 8: Lower bound by number of topics and sample size

5.1.2. Selection of correlation threshold (ρ_0)

As defined in the previous section the correlation threshold sets the limit to which the correlation between topics are considered as meaningful. This determines the amount of connections in the adjacency matrix, affecting the number of newly generated attributes and the topic correlation density. We analyzed how different values of ρ_0 affect the performance of the proposed link prediction approach. The results are discussed below.

Topic correlation density describes the portion of the potential connections in the network of correlated topics that are actual connections. The results show that the maximum density is

reached in the model with K = 50, then the density declines as the number of topics increases. In the neuroscience model the correlation density has a more pronounced decreasing rate.

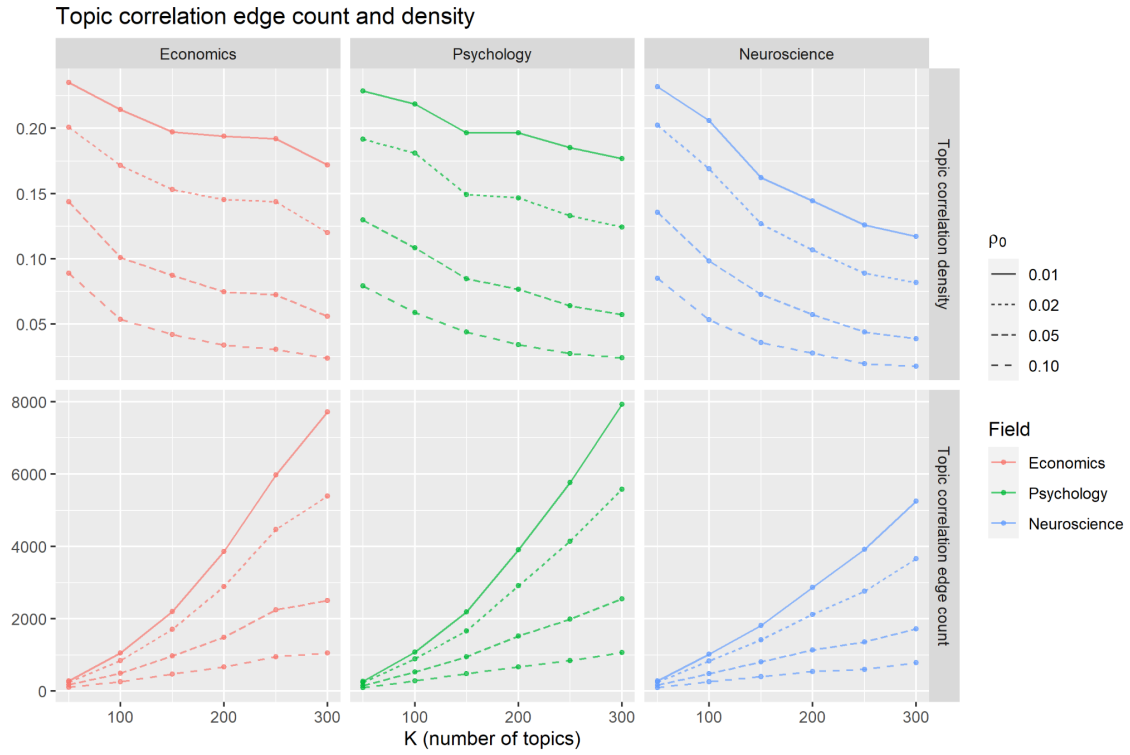


Figure 9: Size and density of the topic network by number of topics and correlation threshold

Topic correlation edge count indicates the number of new attributes generated by the summarizing function. The number of edges in the correlation network grows along with the number of topics however it does slower than the number of potential connections explaining the density decline.

The trade-off between incorporating more topic correlation information and keeping the number of new attributes limited must be taken in consideration when choosing ρ_0. The same criteria is useful to decide the number of topics K. For example: parameter combination K = 100, ρ_0 = 0.01 generates similar number of attributes than K = 200, ρ_0 = 0.05.

We evaluate the link prediction performance of the proposed approach in the two alternative domains, using geometric mean, with K ranging from 50 to 300, and with ρ_0 taking two values 0.01 and 0.05.

The results show that in most cases models with a correlation threshold of 0.01 achieved the highest values of AUC-ROC and AUC-PRC evaluation measures. These results were consistent regardless of the domains and scientific fields with the exception of AUC-ROC in neuroscience

where the model with ρ_0= 0.01 dropped to second place by a small margin. This positive effect of lowering the correlation threshold on performance is strong evidence that including more topic correlations in the node attributes add new information to the model that better represent the link formation process.

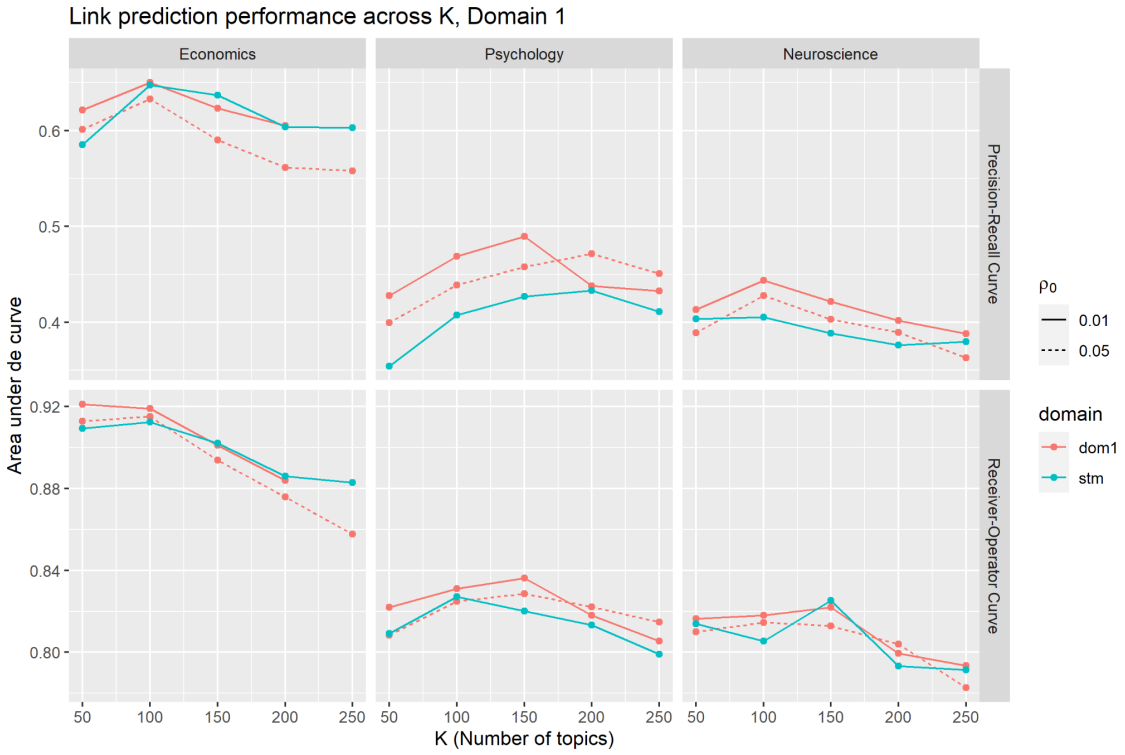


Figure 10: Area under the PR and ROC curves of cosine similarity of domain 1 attributes by number of topics and correlation threshold

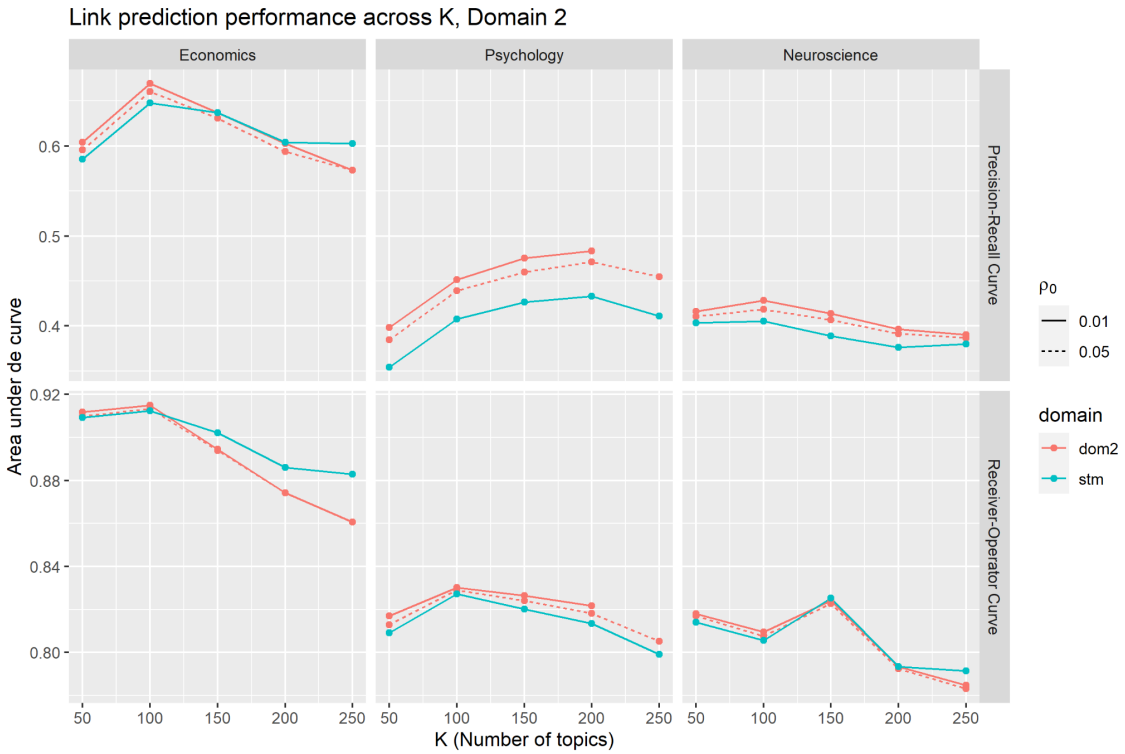


Figure 11: Area under the PR and ROC curves of cosine similarity of domain 2 attributes by number of topics and correlation threshold

5.1.3. Selection of the function domain

Regarding the function domain we evaluate the link prediction performance of both alternatives in models, using geometric mean, with K ranging from 50 to 300, and with ρ_0 fixed at 0.01.

The results of this test are not decisive. There was not a domain that outperformed the other in the same aspect in a consistent way. Psychology was the only field where domain 1 achieved the highest value in both measures AUC-ROC and AUC-PRC, for Neuroscience and Economics the results are mixed.

When analyzing the AUC-PRC performance the results are also mixed, domain 1 achieved the highest value in two fields, while domain 2 ranked first in the field of economics. The same occurs for AUC-ROC performance, domain 2 ranked first only in neuroscience by a small margin.

Although it could not be determined which domain is best for link prediction, what we did find is that in most cases the attributes generated by the proposed approach outperformed those obtained directly from the original structural topic model. And in the cases where STM ranked first it was barely higher than the values achieved by domain 1 and domain 2.

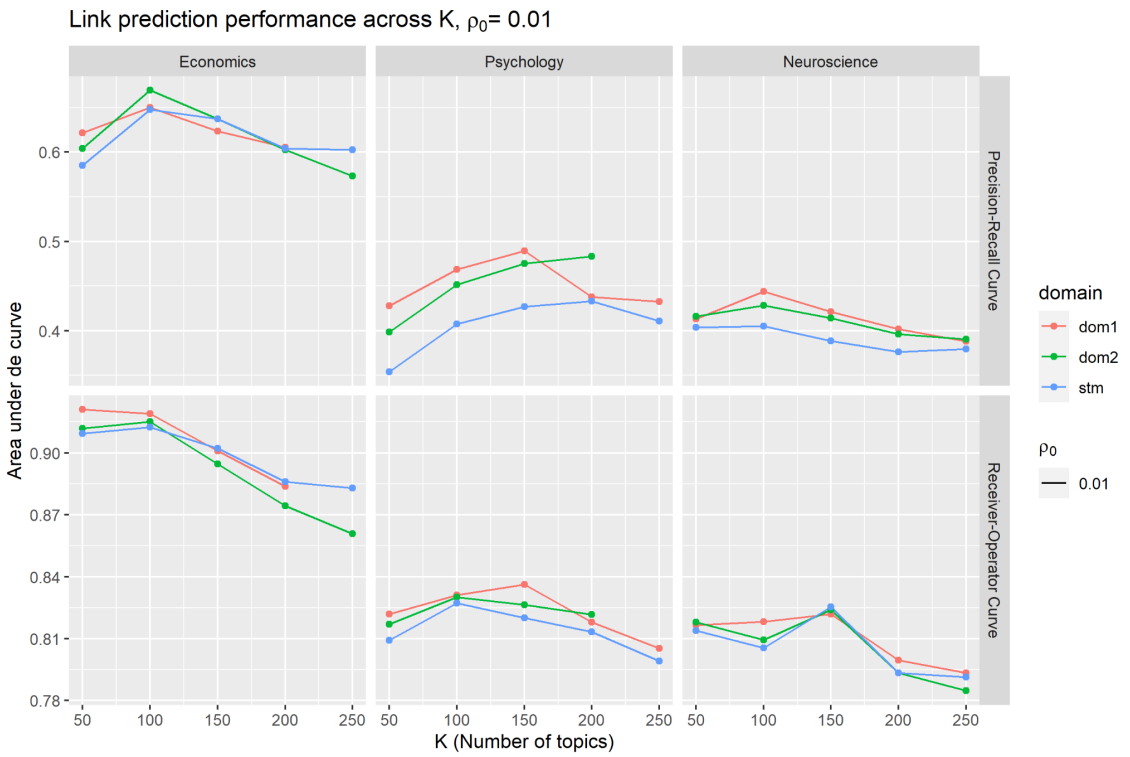


Figure 12: Area under the PR and ROC curves of cosine similarity attributes by number of topics and domain

5.1.4. Selection of summarizing function

For the selection of the summarizing operation we compared the link prediction performance of the three alternatives in a setup with a topic number K = 100 and a correlation threshold of ρ_0 = 0.01. The evaluation was conducted for each domain.
The results show that geometric mean obtained consistently better results in both AUC-ROC and AUC-PRC measures. Only in the field of economics with domain 1 the geometric mean dropped to second place by a small margin.

Table 7: Performance evaluation of attribute summarizing function

Attribute summarizing function		Economics	Psychology	Neuroscience
Domain	Operation	ROC / PRC	ROC / PRC	ROC / PRC
Domain 1	Geometric mean	**0.919**(1) / 0.650	**0.831**(1) / **0.469**(1)	**0.818**(1) / **0.444**(1)
	Arithmetic mean	0.917 / 0.655(2)	0.823 / 0.421	0.812 / 0.409
	Maximum	0.915 / 0.644	0.820 / 0.407	0.810 / 0.401
Domain 2	Geometric mean	0.915(2) / **0.669**(1)	0.830(2) / 0.451(2)	0.809(2) / 0.428(2)
	Arithmetic mean	0.915(2) / 0.652	0.824 / 0.420	0.810 / 0.409
	Maximum	0.915(2) / 0.644	0.822 / 0.409	0.809 / 0.402

5.2. Model evaluation and baseline comparison

The assessment of the link prediction capability of the proposed approach involved analyzing the similarity relation to the network structure, and comparing the performance in the link prediction task to a baseline of text-processing techniques.

5.2.1. Similarity decay across distance

The homophily principle is a basic organizing principle in social networks by which people connect with those who are similar to them (McPherson et al., 2001). This tendency generates assortative mixing patterns in the network structure (Newman, 2003). According to this principle dyadic similarity is strongly influenced by structural factors.

We examined how the proposed attribute similarity of a pair of authors distributes according to their distance in the network. For each distance between 1 and 10 we selected 100 pair of

authors and calculated the cosine similarity using the attributes generated by a topic model with K = 100, correlation threshold of ρ_0 = 0.01, with domain 1 and geometric mean summarization.

The results show a considerable drop in similarity between connected pairs of nodes (distance equals 1) and pairs of nodes at greater distance. From distance 2 forward the similarity continues decreasing at different rates. In economics and psychology the similarity decays at slow rates. In contrast, in the field of neuroscience similarity keeps dropping, reaching its minimum value at distance equals 5. We speculate that differences in the magnitude of the similarity drop could be related to the density and clustering coefficient of the network (Table 2).

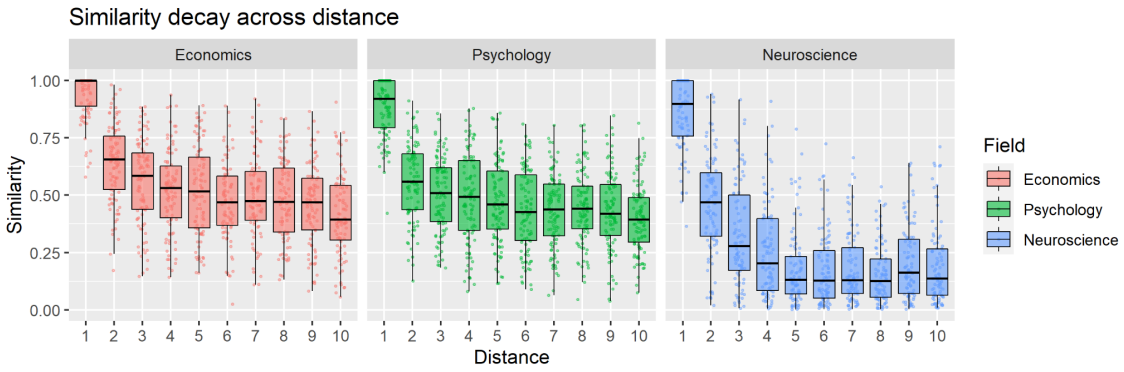


Figure 13: Similarity decay across distance

5.2.2. Link prediction performance

For the purpose of the comparison of the proposed approach to the baseline methods we selected four models obtained from combining the two domains and two values of K, 100 and 200. All four models were obtained with a correlation threshold of ρ_0 = 0.01, and geometric mean summarization.

The experimental results are summarized in Table 8 which present the area under the curve (AUC) of the receiver-operator characteristic curve (ROC) and the precision-recall curve (PRC). At a first glance it can be noticed that our first model (Domain 1, K = 100) achieves the highest AUC-ROC values across all three scientific fields. For the sake of a fair comparison we acknowledge that STM similarity overperform our model at K = 150 as shown in Figure 12. In the fields of economics and psychology our third model (Domain 2, K = 100) reaches the second best AUC-ROC value, while in the field of neuroscience both TFIDF similarity measures tie in second place.

Table 8: Performance evaluation of the proposed approach

Attribute@K	Economics		Psychology		Neuroscience	
AUC	ROC	/ PRC	ROC	/ PRC	ROC	/ PRC
TFIDF@100	0.891	/ 0.659	0.802	/ 0.482	<u>0.813</u>[2]	/ <u>0.459</u>[2]
TFIDF@200	0.893	/ <u>0.668</u>[2]	0.811	/ **0.491**[1]	<u>0.813</u>[2]	/ **0.461**[1]
NMF@100	0.866	/ 0.576	0.790	/ 0.403	0.789	/ 0.362
NMF@200	0.881	/ 0.610	0.789	/ 0.391	0.789	/ 0.391
LDA@100	0.912	/ 0.664	0.804	/ 0.402	0.790	/ 0.412
LDA@200	0.893	/ 0.654	0.802	/ 0.418	0.787	/ 0.404
STM@100	0.912	/ 0.648	0.827	/ 0.408	0.805	/ 0.405
STM@200	0.886	/ 0.604	0.813	/ 0.433	0.793	/ 0.376
W2V@100	0.898	/ 0.658	0.768	/ 0.422	0.793	/ 0.434
W2V@200	0.898	/ 0.656	0.767	/ 0.419	0.795	/ 0.440
GLOVE@100	0.884	/ 0.622	0.773	/ 0.377	0.790	/ 0.404
GLOVE@200	0.806	/ 0.542	0.708	/ 0.336	0.704	/ 0.359

DOM1@100	**0.919**[1] / 0.650	**0.831**[1] / 0.490[2]	**0.818**[1] / 0.444
DOM1@200	0.884 / 0.605	0.829 / 0.403	0.800 / 0.402
DOM2@100	0.915[2] / **0.669**[1]	0.830[2] / 0.451	0.809 / 0.428
DOM2@200	0.874 / 0.603	0.822 / 0.483	0.793 / 0.396

The highest AUC-PRC value in economics was achieved by our third model (Domain 2, K = 100) and the second higher by TFIDF similarity (K = 200). In psychology, TFIDF similarity (K = 200) came out in the first place and our first model (Domain 1, K = 100) in second place. And in neuroscience, TFIDF similarity took both first (K = 200) and second place (K = 100).

Regarding the number of topics, we observed different effects in the text-based similarity performance. Similarity measures based on topic models like LDA, STM and the four proposed models performed better at lower number of topics (K = 100). The results of GloVe were also better at a lower rank (K = 100). In contrast, TFIDF results improved with a higher limit on the number of words for each document (K = 200). Regarding NMF, the results were mixed, it performed better with fewer dimensions (K = 100) in both, economics and neuroscience, but in psychology it did better with more dimensions (K = 200). The similarity based on Word2vec was the least affected measure to changes in the number of dimensions and the effects across fields were also mixed.

5.2.3.Similarity correlations

In order to better understand the differences in performance we analyzed the correlation between the text-based similarity measures. The results of the correlation analysis are presented in Figure 14.

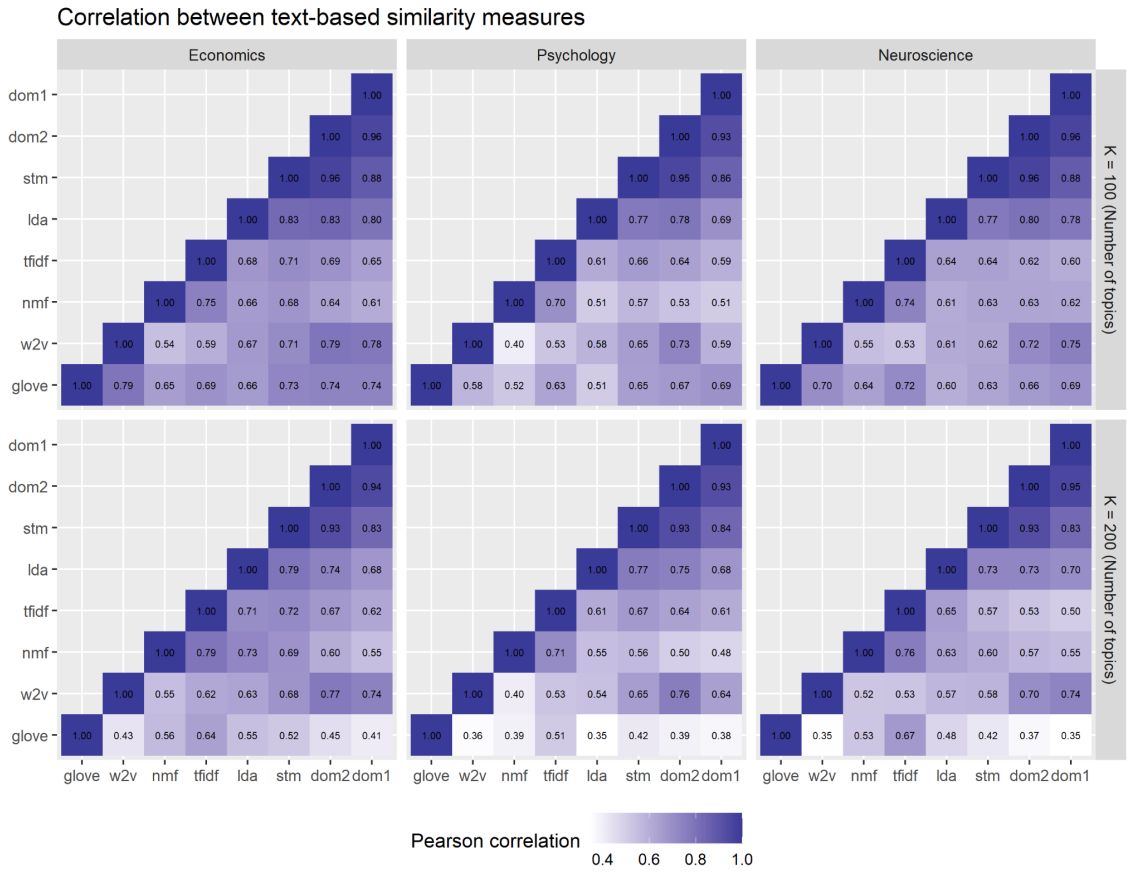


Figure 14: Correlation between text-based similarity measures

The value in each cell is the Pearson's correlation coefficient between the two corresponding similarity vectors used in the link prediction task. This means that each element of a vector is the similarity value for a pair of authors.

The results show strong correlations between the similarity measures based on topic models like LDA, STM and the proposed measures. The measure built upon domain 2 correlates very strongly (0.9 ~ 1.0) to STM and domain 1 measures. This was to be expected since domain 2 attributes are a combination of both. Correlation between domain 1 and STM measures is also strong (0.8 ~ 0.9). The LDA measure was found to be strongly correlated with the STM (0.73 ~ 0.83), domain 2 (0.73 ~ 0.83) and domain 1 (0.68 ~ 0.80) measures.

To our surprise, the correlation values between the TFIDF measure with topic-based similarity measures resulted in the moderate range (0.50 ~ 0.72), in spite of having comparable results to the best performing topic-based measures. In particular, the correlation between TFIDF and domain 1 measures is lower in comparison to the rest of topic-based measures (0.50 ~ 0.65).

Regarding the NMF measure, the strongest and more consistent correlation was found between NMF and the TFIDF measure (0.70 ~ 0.79).

The correlation between Word2vec with the group of measures based on topic models was moderate to strong (0.58 ~ 0.78). Particularly, the strongest correlations occurred between Word2vec and domain 2 measure in the fields of economics (0.77 ~ 0.79) and psychology (0.73 ~ 0.76), and between Word2vec and domain 1 measure in the field of neuroscience (0.75 ~ 0.75).

The rank value had an important influence on the correlation between GloVe with other measures. When the rank was set to K = 100 the correlation between GloVe with all the other measures resulted in the moderate to strong range (0.51 ~ 0.79). When the rank increased to K = 200 the correlation with other measures dropped to the weak to moderate range (0.35 ~ 0.67).

6.Conclusions

In this paper, we have introduced a simple yet effective approach to enrich text-based node attributes for link prediction by incorporating information about correlation between topics. We also tested different summarizing functions and parameter values to find the optimal configuration of the model. In most of the tested configurations the proposed approach outperformed the base topic model.

We found that our approach performs better when using the geometric mean as a summarizing function. This implies that the influence of a pair of correlated topics in the formation of new links is greater when scholars are interested in both topics to some degree, rather than when they are interested in only one of the topics. Also, increasing the correlation threshold showed a reduction in the model performance, indicating that even small correlations between topics carry relevant information.

In general, the inclusion of the original topic distributions in the computed node attributes was found to slightly decrease the performance in both AUC-ROC and AUC-PRC values with a few exceptions.

We explored the assortativity property of the proposed approach by randomly selecting pairs of nodes at different distances in the network. Consistent with the expectations of the homophily principle, we find that similarity decreases significantly with the increase in network distance.

In comparison to the baseline, we found that our approach outperformed traditional text-processing methods in the AUC-ROC value. When considering the AUC-PRC values the proposed approach is positioned among the top ranked methods. Similarity measures based on traditional TFIDF attributes showed performance levels close to those obtained by the proposed approach, even surpassing it in some cases.

6.1.Direction for future research

Our future research comprises various optimization challenges and possible applications. A priority is refining the proposed approach by including covariates in the model estimation. The STM model allows the use of metadata as covariates to estimate topic prevalence and topic content. Although the proposed approach focuses on topic correlations we do not discard the incorporation of documents metadata (e.g. publication year, publication venues, author affiliations, citation count, cited references, etc) as covariates in future research.
Another opportunity for future research is to incorporate supervised learning methods for the combination of different text-based and topology-based similarity measures. This could deepen our knowledge about differences and similarities between the information contributed into authors attributes by topic-correlations and traditionally used text-representation methods.

Regarding the applications, our approach could also be useful to address related research problems like modeling the formation of giant components using cold start link prediction based on text attributes and modeling the formation of interdisciplinary scientific collaboration. In addition, our approach could be useful for developing applications to recommend potential collaboration to scholars based on their publications, or declared research interests.

7.Bibliography

Liben-Nowell, D., & Kleinberg, J. (2007. The link-prediction problem for social networks. Journal of the American society for information science and technology, 58(7, 1019-1031.

Al Hasan, M., & Zaki, M. J. (2011. A survey of link prediction in social networks. In Social network data analytics (pp. 243-275. Springer, Boston, MA.

Wang, P., Xu, B., Wu, Y., & Zhou, X. (2015. Link prediction in social networks: the state-of-the-art. Science China Information Sciences, 58(1, 1-38.

Martínez, V., Berzal, F., & Cubero, J. C. (2016. A survey of link prediction in complex networks. ACM computing surveys (CSUR, 49(4, 1-33.

Pandey, B., Bhanodia, P. K., Khamparia, A., & Pandey, D. K. (2019. A comprehensive survey of edge prediction in social networks: Techniques, parameters and challenges. Expert Systems with Applications, 124, 164-181.

Samad, A., Qadir, M., Nawaz, I., Islam, M. A., & Aleem, M. (2020. A comprehensive survey of link prediction techniques for social network. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, 7(23.

Cai, H., Zheng, V. W., & Chang, K. C. C. (2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering, 30(9, 1616-1637.

Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M. (2006, April. Link prediction using supervised learning. In SDM06: workshop on link analysis, counter-terrorism and security (Vol. 30, pp. 798-805.

Wohlfarth, T., & Ichise, R. (2008, November. Semantic and event-based approach for link prediction. In International Conference on Practical Aspects of Knowledge Management (pp. 50-61. Springer, Berlin, Heidelberg.

Bartal, A., Sasson, E., & Ravid, G. (2009, July. Predicting links in social networks using text mining and sna. In 2009 International conference on advances in social network analysis and mining (pp. 131-136. IEEE.

Liu, Y., Niculescu-Mizil, A., & Gryc, W. (2009, June. Topic-link LDA: joint models of topic and author community. In proceedings of the 26th annual international conference on machine learning (pp. 665-672.

Sachan, M., & Ichise, R. (2010. Using semantic information to improve link prediction results in network datasets. International Journal of Engineering and Technology, 2(4, 334.

Bhattacharyya, P., Garg, A., & Wu, S. F. (2011. Analysis of user keyword similarity in online social networks. Social network analysis and mining, 1(3, 143-158.

Makrehchi, M. (2011, October. Social link recommendation by learning hidden topics. In Proceedings of the fifth ACM conference on Recommender systems (pp. 189-196.

Chuan, P. M., Ali, M., Khang, T. D., & Dey, N. (2018. Link prediction in co-authorship networks based on hybrid content similarity metric. Applied Intelligence, 48(8, 2470-2486.

Rahmaida, R., Saefuddin, A., & Sartono, B. (2019. Predicting Potential Co-Authorship Using Random Forest: Case of Scientific Publications in Indonesian Institute of Sciences. STI Policy and Management Journal, 4(2.

Zhang, Y., Shen, S., & Wu, Z. (2018, August. Improve link prediction accuracy with node attribute similarities. In International Conference on Computer Engineering and Networks (pp. 376-384. Springer, Cham.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Barabási, A. L. (2018). Science of science. Science, 359(6379), eaao0185.

Newman, M. E. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical review E*, *64*(1), 016131.

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. Reviews of modern physics, 74(1), 47.

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research policy*, *34*(10), 1608-1618.

Molontay, R., & Nagy, M. (2021). Twenty years of network science: A bibliographic and co-authorship network analysis. In *Big data and social media analytics* (pp. 1-24). Springer, Cham.

Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, *714*, 1-73.

Abt, H. A. (2007). The future of single-authored papers. *Scientometrics*, *73*(3), 353-358..

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036-1039.

Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of informetrics*, *4*(4), 540-553.

Yuliansyah, H., Othman, Z. A., & Bakar, A. A. (2020). Taxonomy of link prediction for social network analysis: A review. *IEEE Access*, *8*, 183470-183487.

Pham, P., & Do, P. (2019). W-MetaPath2Vec: The topic-driven meta-path-based model for large-scaled content-based heterogeneous information network representation learning. *Expert Systems with Applications*, *123*, 328-344.

Shu, L., Chen, C., Xing, X., Liao, X., & Zheng, Z. (2022. AHNA: Adaptive representation learning for attributed heterogeneous networks. International Journal of Intelligent Systems, 37(2, 1157-1185.

Wang, T., Yuan, W., & Guan, D. (2021, January. Attributed Heterogeneous Network Embedding for Link Prediction. In Pacific Rim Knowledge Acquisition Workshop (pp. 106-119. Springer, Cham.

Xu, S., Yang, C., Shi, C., Fang, Y., Guo, Y., Yang, T., ... & Hu, M. (2021, October. Topic-aware heterogeneous graph neural network for link prediction. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (pp. 2261-2270.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535-74.

Yoon, B., & Magee, C. L. (2018. Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction. Technological Forecasting and Social Change, 132, 105-117.

Hettige, B., Li, Y. F., Wang, W., & Buntine, W. (2020, February. Gaussian embedding of large-scale attributed graphs. In Australasian database conference (pp. 134-146. Springer, Cham.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Solaimannezhad, H., & Fatemi, O. (2017. Representing a Content-based link Prediction Algorithm in Scientific Social Networks. Information Systems & Telecommunication, 146.

Ho, T. K. T., Bui, Q. V., & Bui, M. (2019, December. Co-author relationship prediction in bibliographic network: A new approach using geographic factor and latent topic information. In Proceedings of the Tenth International Symposium on Information and Communication Technology (pp. 69-77.

Wong, S. K. M., Ziarko, W., Raghavan, V. V., & Wong, P. C. (1987). On modeling of information retrieval concepts in vector spaces. *ACM Transactions on Database Systems (TODS)*, *12*(2), 299-321.

Fellbaum, C. (1998). A semantic network of english: the mother of all WordNets. In *EuroWordNet: A multilingual database with lexical semantic networks* (pp. 137-148). Springer, Dordrecht.

Zarrinkalam, F., Fani, H., Bagheri, E., & Kahani, M. (2016, March. Inferring implicit topical interests on twitter. In European Conference on Information Retrieval (pp. 479-491. Springer, Cham.

Witten, I. H., & Milne, D. N. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links.

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013, December). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation* (Vol. 4, pp. 1-20).

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The annals of applied statistics*, *1*(1), 17-35.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, *91*, 1-40.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833.

Baek, J. W., & Chung, K. Y. (2021. Multimedia recommendation using Word2Vec-based social relationship mining. Multimedia Tools and Applications, 80(26, 34499-34515.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, *1*(4), 309-317.

Starrfield, S., Truran, J. W., Sparks, W. M., & Kutter, G. S. (1972). CNO abundances and hydrodynamic models of the nova outburst. *The Astrophysical Journal*, *176*, 169.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788-791

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009, June). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105-1112).

Taddy, M. (2012, March). On estimation and selection for topic models. In Artificial Intelligence and Statistics (pp. 1184-1193). PMLR.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272).

Tzikas, D. G., Likas, A. C., & Galatsanos, N. P. (2008). The variational approximation for Bayesian inference. IEEE Signal Processing Magazine, 25(6), 131-146.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, 415-444.

Newman, M. E. (2003). Mixing patterns in networks. *Physical review E*, *67*(2), 026126.