

Semantic Similarity Measures for Topological Link Prediction

Giulio Biondi[1] and Valentina Franzoni[2(✉)]

[1] Department of Mathematics of Computer Science, University of Florence, 50134 Florence, Italy
giulio.biondi@unifi.it
[2] Department of Mathematics of Computer Science, University of Perugia, 06123 Perugia, Italy
valentina.franzoni@dmi.unipg.it

Abstract. The semantic approach to data linked in social networks uses information extracted from node attributes to quantify the similarity between nodes. In contrast, the topological approach exploits the structural information of the network, e.g., nodes degree, paths, neighbourhood breadth. For a long time, such approaches have been considered substantially separated. In recent years, following the widespread of social media, an increasing focus has been dedicated to understanding how complex networks develop, following the human phenomena they represent, considering both the meaning of the node and the links structure and distribution. The link prediction problem, aiming at predicting how networks evolve in terms of connections between entities, is suitable to apply semantic similarity measures to a topological domain. In this paper, we introduce a novel topological formulation of semantic measures, e.g., NGD, PMI, Confidence, in a unifying framework for link prediction in social graphs, providing results of systematic experiments. We validate the approach discussing the prediction capability on widely accepted data sets, comparing the performance of the topological formulation of semantic measures to the conventional metrics generally used in literature.

Keywords: Unified view · Complex networks · Graph-based link prediction · Structural link prediction · Ranking-based approach

1 Introduction

For a long time, the semantic and topological approaches to the similarity in complex networks have been considered substantially separated. The semantic approach uses information extracted from node attributes to quantify the similarity between nodes, while the topological approach exploits the structural information of the network, e.g., nodes degree, paths, neighbourhood [4,7]. A set of techniques have been proposed in the literature for the link prediction, based on topological or semantic similarity. The idea behind both approaches is

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12253, pp. 132–142, 2020.
https://doi.org/10.1007/978-3-030-58814-4_10

that similar nodes will develop a link in the future. Such similarity is structural in the topological approach, feature-based in the semantic approach, but can be applied to the same data. The link prediction problem, aiming at predicting how networks evolve in terms of connections between entities, is usually addressed with a topological approach. Still, we consider also suitable to apply the semantic approach to a topological domain. Formally, the link prediction problem consists in predicting the evolution of a given network at time t in terms of new links that will appear in the network in a future time t'. Forecasting a new link can be valuable for recommending, advertising or generating support, e.g., analysing social networks. E.g., in co-authorship networks, nodes represent authors, links represent co-authorship and prediction forecasts collaborations. Although the two approaches have been integrated into hybrid similarity frameworks, a dichotomy still exists, and the most significant interest in link prediction literature has been devoted to the topological approach, easier to exploit, and domain-independent. The key idea is that the web-based similarity measures use statistics from a search engine which can be interpreted in terms of the structure of the Web (i.e., of the indexing network of search engines). Therefore, such measures indirectly embed a measure of documents' neighbourhood topology. This work shows how semantic similarity measures (i.e., average confidence, PMI, NGD, X-squared) can be reformulated in a topological view and used in an associated graph for link prediction, starting from the concept that a substantial equivalence exists between the two approaches to the Link Prediction problem [9,12]. The advantages of modelling both classes of similarities in a single comprehensive model are apparent: a unified model provides the opportunity to use semantic measures in a topological framework and vice versa. In the first section, we introduce the link prediction problem in our unified framework. In the second section, we present a mapping of the semantic measures to the topological domain. In the third and fourth sections, we experimentally validate the proposed approach, discussing the prediction capability on widely accepted data sets, comparing the performance of the topological formulation of semantic measures to the conventional metrics generally used in literature. In the fifth section, conclusions are drawn.

2 Mapping Semantic Measures in a Topological Domain

2.1 Previous Research

In previous works [9,12] we studied the feasibility of a bidirectional mapping from topological measures to semantic formulations and vice versa and demonstrated the correctness of the unified approach. The comparison has been conducted on web-based semantic proximity measures, emphasizing their ability to express topological concepts by exploiting the capabilities of search engines. In a preliminary phase we performed an in-depth analysis of such measures. We extended the comparison to other proximity measures that can be applied to a web-based domain even if they are not created for that context, obtaining promising results on Pointwise mutual information (PMI) [26]; Normalized Google Distance (NGD) [6]; PMING distance [8], which is a combination of the

previous sensibly improving measuring performances in clustering; Confidence; Average Confidence, seen as a combination of directional Confidence either way; Chi-square. Such measures have been tested on image classification, using image metadata and captions, and on text-based emotion recognition. A proximity model for randomized Heuristic Semantic Search (HSW) [14] has been experimented in several variations to guide the navigation through network links, where semantic measures can be used as heuristics to choose the step-by-step path instead of a random or a rule-based selection from candidate nodes, ranked using an aggregation function on the results of several random tournaments. Previous contributions also include the extension of the context-based Image Semantic Similarity [10] to a new class of set similarity measure to be used on social communities, using different combination schemes. The study of distances based on random walks, which are robust and computationally efficient for context extraction in large unknown graphs, lead to exciting results in terms of topological goals, e.g., convergence, path length, and minimal solution. Such models and techniques have been systematically tested on benchmark data sets, e.g., social networks, bibliographic repositories, recommender systems, and semantic associations between terms and emotional tags. The previous work laid the foundation for investigating algorithms for ranking graph nodes for many applications, including link prediction and reconstruction.

2.2 Topological Formulation of Semantic Similarity

We start considering as information source a search engine S, providing statistics about the occurrence, co-occurrence and probability of terms in a context [13], i.e. the frequency of single terms and terms pairs, used in web-based semantic proximity measures. Consider the terms-to-documents graph $G_S = (V, E)$ indexed by the search engine. Typify G_S as a bipartite graph where the set $V = T \cup D$ can be partitioned in two subsets. Vertices in T represent indexed terms, vertices in D represent the indexed documents. The edges $(u, v) \in E$ are the occurrences of terms in documents, i.e. (u, v) is the relation "u occurring in document v". For each edge in G_S, the property $\forall (u, v) \in E, u \in T \wedge v \in D$ is hold, i.e. G_S is bipartite.

If we consider the sample graph G_S shown in Fig. 1, it is apparent that all the information required by the semantic similarity measures can be defined in terms of neighbourhood and graph parameters of G_S, with the constraint of knowing the considered entities, i.e., the *frequency/occurrence* of a term u, the *co-occurrence* of u and v, and the corresponding *probabilities*

$$f(u) = |\Gamma(u)| \tag{1}$$

and

$$f(u, v) = |\Gamma(u) \cap \Gamma(v)|. \tag{2}$$

In such formulation $\Gamma(u)$ is the set of documents that contain a term u, and a document d is a common neighbour of the terms u and v, $d \in \Gamma(u) \cap \Gamma(v)$, if d contains both terms. Conversely, $\Gamma(d)$ is the set of terms contained in a given document d and $t \in \Gamma(d_1) \cap \Gamma(d_2)$, if t occurs in both d_1 and d_2.

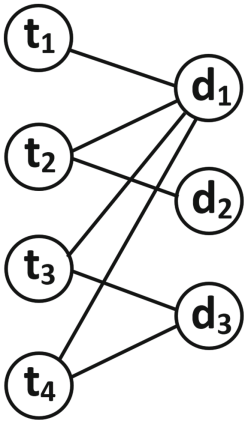

Fig. 1. An example of terms-to-documents graph

Topological Confidence. The *Confidence index (CF)* [2] formulation is straightforward in terms of topology:

$$similarity(u,v)_{TConf} = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u)|} \tag{3}$$

The original Confidence index *TConf* is not symmetric; thus, we define the Average Topological Confidence [13]:

$$similarity(u,v)_{AvgTConf} = \frac{1}{2}\left(\frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u)|} + \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(v)|}\right) \tag{4}$$

Topological Pointwise Mutual Index. The Pointwise Mutual Index (PMI) [5] can be expressed topologically as:

$$similarity(u,v)_{TPMI} = \log_2\left(|V| \cdot \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u)| \cdot |\Gamma(w)|}\right) \tag{5}$$

Topological Normalized Google Distance. The Normalized Google Distance topological formulation results in [6]:

$$similarity(u,v)_{TNGD} = \frac{\max(\log|\Gamma(x)|, \log|\Gamma(y)|) - \log(|\Gamma(x) \cap \Gamma(y)|)}{\log M - \min(\log|\Gamma(x)|, \log|\Gamma(y)|)} \tag{6}$$

Chi-squared Coefficient χ^2. The Chi-squared Coefficient χ^2 [20] can be adapted as follows:

$$similarity(u,v)_{T\chi^2} = \frac{|\Gamma(u) \cap \Gamma(v)| * (|V| - |\Gamma(u) \cup \Gamma(v)|) - |\Gamma(u) * \Gamma(v)|}{(|\Gamma(u) \cap \Gamma(v)| + |\Gamma(u)|)(|\Gamma(u) \cap \Gamma(v)| + |\Gamma(v)|)} * \frac{|V|}{(|\Gamma(u)| + |V| - |\Gamma(u) \cup \Gamma(v)|)(|\Gamma(v)| + |V| - |\Gamma(u) \cup \Gamma(v)|)} \tag{7}$$

Web-based similarity measures such as NGD are explicitly designed to count the frequencies with which terms occur in a corpus, e.g. the documents indexed in a search engine; such corpora indeed represent a bipartite graph, where document nodes are connected to word nodes, and vice versa. In some cases we will not be able to know the real value of M, representing the total number of indexed documents. In this case, we can set the parameter as a random number greater than any occurrence of terms or pairs (see [6,13]), $M > |V|$, because it simply serves as a normalisation value.

3 Design of the Experiments

3.1 Data Sets

This section contains an overview of the data sets, which represent social networks, in particular email correspondence and co-authorship. For each of them, the size after preprocessing, in terms of nodes and edges, is shown in Table 1. The size of the data sets analyzed varies from a few hundreds of nodes (*radoslaw-email, email-EU-core*) to thousands (netscience, ego-Facebook, ca-GrQc, feather-lastfm-social) up to tens of thousands (*CA-HepPh, CA-HepTh, CA-AstroPh*); all of them can be considered sparse with respect to the number of edges. The data sets come from three widely accepted sources, i.e. *Stanford Large Network Dataset Collection* [18], *The Koblenz Network Collection* [16] and *The Network Repository* [24].

Table 1. Data sets statistics

Data set	$\|V\|$	$\|E\|$	Reference
Netscience	1461	2742	[23]
CA-GrQc	5241	14484	[17]
CA-HepPh	12006	118489	[17]
CA-HepTh	9875	25973	[17]
CA-AstroPh	18771	198050	[17]
ia-radoslaw-email	167	3250	[22]
ego-facebook	4039	88234	[21]
email-eu-core	986	16864	[17,28]
feather-lastfm-social	7624	27806	[25]

3.2 Preprocessing

Before the evaluation process, data sets are cleaned and randomised; the following elements are removed since they are not exploited in our approach:

- multiple edges between the same pair of nodes;
- self-loops;
- additional information associated with edges, e.g. text attributes.

3.3 *k*-fold Validation and the 0-Tail Issue

A k-fold validation is performed on each data set, choosing $k = 10$. This setup implies examining particular cases that might appear. The most relevant issue to consider is that entities can appear as 0-degree nodes because all the incident edges have been moved in $E_{TE}(k)$. Therefore, the calculated value for pairs including such nodes may be incorrect for some measures, while others are not affected. In particular:

- In Average Confidence, the degrees of nodes appear as denominators in a fraction; thus, if at least one of them is zero, the case in which the predicted value is infinite is replaced with 0.
- In Jaccard, if both nodes have degree 0, the predicted value is $\frac{0}{0}$, which evaluates to NaN and is replaced with 0.
- In PMI, it is enough to have a 0-degree node to get a $\frac{0}{0}$ value, again NaN, replaced with $-\infty$.
- In NGD, in case of division by zero, the value is replaced with $+\infty$.
- In χ^2, NaN values are replaced with 0.

This solution is correct from the logical point of view, because calculating neighbourhood-based similarity measures between two nodes u and v when one, or even both of them, have degree 0, i.e. have no neighbourhood, leads to a 0-tail [3]. The table containing the prediction values is sorted to produce the ranking induced by similarity measures; in case of distance measures, e.g. NGD, it is sufficient to reverse the rank order prior to the evaluation.

3.4 Evaluation Metrics and Approach-Dependent Considerations

In our approach, Link Prediction can be seen as a Binary Classification (BC) problem because each edge in the E_{POT} set either belongs or not to E_{TE}, used to test the prediction. The following quantities are defined to evaluate the performances of our classification model in the BC view:

- $TP(TN)$: number of positive (negative) instances correctly classified
- $FP(FN)$: number of misclassified negative (positive) instances
- $P_{real} = TP + FN$: number of positive instances
- $N_{real} = TN + FP$: number of negative instances
- $P_{alg} = TP + FP$: number of instances classified as positive
- $N_{alg} = TN + FN$: number of instances classified as negative

The length of P_{alg} and N_{alg} depends on a threshold $\tau = |E_{TE}|$ [19]. Links appearing in the top $|E_{TE}|$ positions in the rank are thus classified as positive, i.e. likely to appear at time $t + 1$. The fraction of links also appearing in $|E_{TE}|$ constitutes the TP set, while the remaining links correspond to the FP set. On the other hand, edges ranked lower than $|E_{TE}|$ are classified as negative, and part of them will not be included in the $|E_{TE}|$, forming the TN set. The remainder composes the FN set. The traditional evaluation metrics for classification are *Accuracy*, *Precision*, *Recall* and *F1*. It is worth noticing that the

choice of our threshold $\tau = |E_{TE}|$ induces additional constraints. In particular, the number of FP instances is equal to FN; therefore, in our approach to the link prediction problem, Precision, Recall and F1 all assume the same values and can be considered equal.

Further considerations are required regarding Accuracy. Our algorithm evaluates the similarity measures for each edge in the set $E_{POT} = E_{FULL} \setminus E_{TR}$, which cardinality is $|E_{POT}| = |E_{FULL}| - |E_{TR}| = |V| \cdot (|V| - 1)) - |E_{TR}|$. The cardinality of the set E_{TE} is an upper bound for the TP, FP and FN, usually small if compared to $|E_{POT}|$. Since $|E_{POT}| = TP + TN + FP + FN$, it follows that $TN \gg TP + FP + FN$. Recalling that accuracy is defined as $ACC = \frac{TP+TN}{TP+FP+TN+FN}$, and given that TN represents the highest quota of the total edges, its value will always be ≈ 1, thus losing its significance. Since Precision considers the top $|E_{TE}|$ edges in the rank induced by similarity measures and counts the positive hits to evaluate performances, in some cases it could produce excessively low scores, not reflecting the real performance of the ranking as a whole. Due to our constraints and considerations, we decided to focus on an additional evaluation metric. Considering the position of the FN values in the rank, to detect whether most of them lay close to the τ ranking position (i.e., the frontier) or, instead, towards the bottom, it is useful to plot the Receiver Operating Characteristic (ROC) curve and calculate the Area Under the Curve (AUC) [27] value. A random predictor with uniform distribution returns AUC values close to 0.5; a higher value denotes a ranking where the positive edges are generally located in the upper section, and 1 indicates perfect rank. Since the accuracy is not significant in our case, we focus on precision and AUC in order to have more general, inter-dataset, key performance indicators.

4 Experimental Results

In the experiments, the mapped measures have been compared to conventional, widely used topological measures for Link Prediction, i.e. Common Neighbours [19], Jaccard [15], Adamic Adar [1] and Resource Allocation [29]. In Tables 2 and 3 the results are averaged on 10 folds; in bold, the best mapped semantic measure for each data set.

Results show that the performance of the mapped semantic measures is generally comparable to topological measures. The χ^2 index delivers highest Precision among the mapped measures in five out of nine experiments, with the remainder equally divided between Average Confidence and NGD. The reason for the higher Precision of AA and RA lies in the information about the neighbourhood up to a distance of 2 links traversal. CN and Jaccard, as well as our mapped measures, exploit only the direct neighbourhood of nodes, and their performance can be therefore compared. The exception is PMI [11], which is not a suitable correlation metric when only a few observations are provided. It is not surprising that the worst performances of semantic-derived measures are obtained with the smallest data sets, where node degree is in the order of few tens or hundreds.

Table 2. Average precision over 10 fold

	CN	Jaccard	AA	RA	Random	AvgConf	NGD	PMI	X2
netscience	0.4478	0.4967	0.6634	0.6823	0.0000	0.4847	**0.4967**	0.1948	0.4916
CA-GrQc	0.3594	0.3521	0.4873	0.5472	0.0001	0.3234	0.3170	0.0554	**0.3368**
CA-HepPh	0.5084	0.5654	0.5894	0.7790	0.0002	0.5410	0.4594	0.0359	**0.5744**
CA-HepTh	0.2179	0.1333	0.3399	0.3361	0.0000	**0.1274**	0.1262	0.0461	0.1231
CA-AstroPh	0.3813	0.3389	0.4490	0.6309	0.0001	0.3213	0.2637	0.0433	**0.3465**
ia-radoslaw	0.4058	0.2634	0.4138	0.4237	0.0320	**0.2748**	0.0308	0.0003	0.0889
Ego-facebook	0.3121	0.3238	0.3273	0.4457	0.0010	0.2767	0.2342	0.0038	**0.3203**
Email-eu-core	0.1995	0.2001	0.2230	0.2654	0.0032	0.0656	0.0938	0.0020	**0.1656**
feather-lastfm	0.1380	0.0074	0.1646	0.1595	0.0001	0.0050	**0.0083**	0.0009	0.0061

Table 3. Average AUC over 10 fold

	CN	Jaccard	AA	RA	Random	AvgConf	NGD	PMI	X2
netscience	0.9335	0.9336	0.9338	0.9338	0.4960	**0.9337**	0.9335	0.9332	0.8819
CA-GrQc	0.9218	0.9218	0.9221	0.9221	0.5019	**0.9219**	0.9218	0.9213	0.8942
CA-HepPh	0.9806	0.9816	0.9815	0.9821	0.5011	**0.9815**	**0.9815**	0.9794	0.9773
CA-HepTh	0.9019	0.9019	0.9022	0.9022	0.4998	**0.9019**	**0.9019**	0.9017	0.8856
CA-AstroPh	0.9885	0.9888	0.9892	0.9895	0.4993	**0.9887**	0.9883	0.9868	0.9861
ia-radoslaw	0.9146	0.8619	0.9167	0.9205	0.5002	**0.8457**	0.6606	0.4620	0.7631
Ego-facebook	0.9925	0.9907	0.9936	0.9949	0.4985	**0.9910**	0.9866	0.9588	0.9905
Email-eu-core	0.9390	0.9298	0.9440	0.9489	0.5002	**0.9203**	0.8918	0.8498	0.9126
feather-lastfm	0.8474	0.8465	0.8478	0.8478	0.4953	0.8463	0.8460	0.8450	**0.8586**

Table 4. 10-fold precision, Netscience data set

Fold	CN	Jaccard	AA	RA	Random	AvgConf	NGD	PMI	X2
0	0.4255	0.5018	0.6545	0.68	0	0.4945	0.5055	0.2109	0.5055
1	0.4836	0.4873	0.6509	0.6836	0	0.4655	0.4873	0.1673	0.48
2	0.4891	0.5547	0.7117	0.7372	0	0.5438	0.5657	0.1825	0.5438
3	0.4197	0.4416	0.6241	0.6606	0	0.4343	0.4453	0.1934	0.438
4	0.4161	0.5219	0.6569	0.6788	0	0.5109	0.5109	0.2226	0.5073
5	0.4489	0.4599	0.6606	0.6715	0	0.4635	0.4489	0.1861	0.4526
6	0.4562	0.4927	0.6679	0.6715	0	0.4891	0.4964	0.208	0.4964
7	0.4453	0.4964	0.6642	0.6752	0	0.4781	0.5	0.1825	0.4964
8	0.4526	0.5037	0.6715	0.6788	0	0.4818	0.5037	0.2007	0.4927
9	0.4416	0.5073	0.6715	0.6861	0	0.4854	0.5037	0.1934	0.5037

Regarding the AUC, the best performances are obtained with Average Confidence, which is also comparable to topological alternatives. In all but few cases, the mapped measures achieve consistent results on all the data sets.

In Table 4, the results of the 10-fold validation process are provided for the Netscience data set. It is apparent that all the measures show a coherent behaviour across all the folds, with only a few exceptions; such behaviour is shared across all the experiments.

5 Conclusion

In this paper, we introduce a new framework for the Link Prediction problem, aiming at giving a unified view of the two currently investigated approaches, topological and semantic. Our framework allows the use of semantic proximity measures to perform Link Prediction in a topological domain by remapping them according to a formal rule. Experiments have been held on data sets of different sizes, spanning from hundreds to tens of thousands of nodes. Different types of networks are represented, e.g., collaboration and e-mail, to evaluate the performance of semantic measures, i.e. Average Confidence, PMI, NGD, and X2, in a topological context. As a general conclusion, our experimental evidence shows that Average Confidence is comparable to classical topological ranking measures in terms of AUC; while χ^2 returns better Precision. We also observed that, in agreement with the literature, the applicability of measures derived from web-based semantics improves as the size of the network increases.

References

1. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Soc. Netw. **25**(3), 211–230 (2003). https://doi.org/10.1016/S0378-8733(03)00009-1
2. Agrawal, R., Imieliundefinedski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD 1993, pp. 207–216. Association for Computing Machinery, New York (1993). https://doi.org/10.1145/170035.170072
3. Chiancone, A., Franzoni, V., Li, Y., Markov, K., Milani, A.: Leveraging zero tail in neighbourhood for link prediction, pp. 135–139 (2016). https://doi.org/10.1109/WI-IAT.2015.129
4. Chiancone, A., Franzoni, V., Niyogi, R., Milani, A.: Improving link ranking quality by quasi-common neighbourhood, pp. 21–26 (2015). https://doi.org/10.1109/ICCSA.2015.19
5. Church, K.W., Hanks, P.: Word association noms, mutual information, and lexicography. In: Proceedings of the 27th Annual Conference of the Association for Computational Linguistics, vol. 16, no. 1, pp. 22–29 (1989). https://doi.org/10.3115/981623.981633
6. Cilibrasi, R., Vitanyi, P.: The google similarity distance, arxiv.org or clustering by compression. IEEE J. Trans. Inf. Theory **51**(4), 1523–1545 (2004)

7. Franzoni, V., Chiancone, A., Milani, A.: A multistrain bacterial diffusion model for link prediction. Int. J. Pattern Recogn. Artif. Intell. **31**(11) (2017). https://doi.org/10.1142/S0218001417590248
8. Franzoni, V., Milani, A.: PMING distance: a collaborative semantic proximity measure, vol. 2, pp. 442–449 (2012). https://doi.org/10.1109/WI-IAT.2012.226
9. Franzoni, V., Milani, A.: Structural and semantic proximity in information networks. In: Gervasi, O., et al. (eds.) ICCSA 2017. LNCS, vol. 10404, pp. 651–666. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62392-4_47
10. Franzoni, V., Milani, A., Pallottelli, S., Leung, C., Li, Y.: Context-based image semantic similarity, pp. 1280–1284 (2016). https://doi.org/10.1109/FSKD.2015.7382127
11. Franzoni, V.: Misure di prossimità semantica per il Web. Master's thesis (2012)
12. Franzoni, V.: A unifiying approach to semantic and topological similarity in information networks. Ph.D. thesis (2017)
13. Franzoni, V., Milani, A.: PMING distance: a collaborative semantic proximity measure. In: Proceedings - 2012 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2012, vol. 2, pp. 442–449 (2012). https://doi.org/10.1109/WI-IAT.2012.226
14. Franzoni, V., Milani, A.: Heuristic semantic walk. In: Murgante, B., et al. (eds.) ICCSA 2013. LNCS, vol. 7974, pp. 643–656. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39649-6_46
15. Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences Naturelles **37**(JANUARY 1901), 547–579 (1901). https://doi.org/10.5169/seals-266450
16. Kunegis, J.: Konect: the koblenz network collection. In: Proceedings of the 22nd International Conference on World Wide Web, WWW 2013, Companion, pp. 1343–1350. Association for Computing Machinery, New York (2013). https://doi.org/10.1145/2487788.2488173
17. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: densification and shrinking diameters. ACM Trans. Knowl. Discov. Data **1**(1), 2-es (2007). https://doi.org/10.1145/1217299.1217301
18. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford large network dataset collection, June 2014. http://snap.stanford.edu/data
19. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inform. Sci. Technol. **58**(7), 1019–1031 (2007). https://doi.org/10.1002/asi.20591
20. Manning, C.D., Schütze, H., Weikurn, G.: Foundations of statistical natural language processing. SIGMOD Rec. (2002). https://doi.org/10.1145/601858.601867
21. McAuley, J., Leskovec, J.: Learning to discover social circles in ego networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, NIPS 2012, vol. 1, pp. 539–547. Curran Associates Inc., Red Hook (2012)
22. Michalski, R., Palus, S., Kazienko, P.: Matching organizational structure and social network extracted from email communication. In: Abramowicz, W. (ed.) BIS 2011. LNBIP, vol. 87, pp. 197–206. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21863-7_17
23. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices, May 2006. https://doi.org/10.1103/PhysRevE.74.036104
24. Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: AAAI (2015). http://networkrepository.com
25. Rozemberczki, B., Sarkar, R.: Characteristic functions on graphs: birds of a feather, from statistical descriptors to parametric models (2020)

26. Turney, P.D.: Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: De Raedt, L., Flach, P. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44795-4_42
27. Yang, Y., Lichtenwalter, R.N., Chawla, N.V.: Evaluating link prediction methods. Knowl. Inf. Syst. **45**(3), 751–782 (2014). https://doi.org/10.1007/s10115-014-0789-0
28. Yin, H., Benson, A.R., Leskovec, J., Gleich, D.F.: Local higher-order graph clustering. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2017, pp. 555–564. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3097983.3098069
29. Zhou, T., Lu, L., Zhang, Y.C.: Predicting missing links via local information. Eur. Phys. J. B **71**(4), 623–630 (2009). https://doi.org/10.1140/epjb/e2009-00335-8