Semantics of User Interaction in Social Media

Folke Mitzlaff, Martin Atzmueller, Gerd Stumme[1], and Andreas Hotho[2]

[1] Knowledge and Data Engineering Group, University of Kassel,
Wilhelmshöher Allee 73, D-34121 Kassel, Germany,
{mitzlaff,atzmueller,stumme}@cs.uni-kassel.de
[2] Data Mining and Information Retrieval Group, University of Würzburg,
Am Hubland, D-97074 Würzburg, Germany
hotho@informatik.uni-wuerzburg.de

Abstract. In ubiquitous and social web applications, there are different user traces, for example, produced explicitly by "tweeting" via twitter or implicitly, when the corresponding activities are logged within the application's internal databases and log files.

For each of these systems, the sets of user interactions can be mapped to a network, with links between users according to their observed interactions. This gives rise to a number of questions: Are these networks independent, do they give rise to a notion of user relatedness, is there an intuitively defined relation among users?

In this paper, we analyze correlations among different interaction networks among users within different systems. To address the questions of interrelationship between different networks, we collect for every user certain external properties which are independent of the given network structure. Based on these properties, we then calculate semantically grounded reference relations among users and present a framework for capturing semantics of user relations. The experiments are performed using different interaction networks from the twitter, flickr and BibSonomy systems.

1 Introduction

With the increasing availability of mobile internet connections, social applications are ubiquitously integrated in the user's daily life. By interacting with such systems, the user is leaving traces within the different databases and log files, e. g., by updating the current status via twitter or chatting with social acquaintances via facebook. Ultimately, each type of such traces gives rise to a corresponding network of user relatedness, where users are connected if they interacted either explicitly (e. g., by establishing a "friendship" link within in an online social network) or implicitly (e. g., by visiting a user's profile page). We consider a link within such a network as evidence for user relatedness and call it accordingly *evidence network* or *interaction network*.

These interaction networks are of large interest for many applications, such as recommending contacts in online social networks or for identifying groups of related users [19]. Nevertheless, it is not clear, whether every such interaction network captures meaningful notions of relatedness and what the semantics of different aggregation levels really are.

As multifaceted humans are, as many reasons for individuals being related exists. Ultimately, it is therefore not possible to judge whether an interaction network is "meaningful" or not. Nevertheless, certain networks are more probable than others and give rise to more traceable notions of relatedness. In this paper, we compare different networks based on external notions of relationship between users, especially geographical proximity. We argue that the geographical proximity of two users is a prior for other important notions of relationship, such as common language and cultural background. We therefore use such an external measure of relatedness as a proxy for semantically grounded relationship among users.

This paper proposes an experimental methodology for assessing the semantics of evidence networks and similarity metrics therein. The contributions of the paper can be summarized as follows: The presented methodology is applied to a broad range of evidence networks obtained from twitter, flickr and BibSonomy as well as different similarity metrics for calculating similarity of nodes within a network. The obtained results thus yield a *semantic grounding* of evidence networks and similarity metrics, which are merely based on structural properties of the networks. Furthermore, we consider both established reference sources such as tagging data, as well as geographical locational data as a proxy for semantic relatednesss. Finally, the collected analysis results, and especially the proposed methodology can serve as a foundation for further applications, such as *community mining* and *link prediction* tasks.

The remainder of this paper is structured as follows: First, we discuss related work in Section 2. In Section 3 we present all considered network data. Next, Section 4 introduces the proposed methodology for assessing semantics of user relatedness in evidence networks and presents according experimental results obtained from the considered networks. In Section 5 we summarize the results and point at future directions of research and applications for the results presented in this paper.

2 Related Work

The present paper tackles the problem of grounding the semantics of user relatedness, induced by online social networks: It combines approaches from the field of *social network analysis* and methods for measuring *distributional similarity*. Furthermore, it also considers *geospatial analysis* in online social media. Notably, the present work is based on [21], where semantics of Wikipedia based co-occurrence networks of named entities are analyzed with respect to category assignments to corresponding articles in Wiktionary as well as corresponding geo-location.

The field of distributional similarity and semantic relatedness has attracted a lot of attention in literature during the past decades (see [3] for a review). Several statistical measures for assessing the similarity of words are proposed, e. g., in [4, 6, 11, 13, 25], especially in the context of social bookmarking systems [2].

The task of calculating similarity between individuals within a social network is closely related to the *link prediction* and *user recommendation* tasks, where "missing" links are predicted based on the network structure. In the context of social networks, the task of predicting (future) links is especially relevant for online social networks, where social interaction is significantly stimulated by suggesting people as contacts which the user might know. From a methodological point of view, most approaches build on different similarity metrics on pairs of nodes within weighted or unweighted graphs [7, 12, 16, 17]. A good comparative evaluation of different similarity metrics is presented in [15]. In [26] a topic sensitive user ranking in the context of online social media is proposed.

The analysis of online social media, the interrelations of the involved actors, and the involved geospatial extents have attracted a lot of attention during the last decades, especially for the microblogging system twitter. A thorough analysis of fundamental network properties and interaction patterns in twitter can be found in [10].

An analysis of a location-based social network with respect to the user attributes is investigated in [14]. Interdependencies of social links and geospatial proximity are investigated in [8, 9, 18, 23, 24], especially concerning the correlation of the probability of friendship links and the geographic distance of the corresponding users. The impact of subgroups of users and communities is analyzed in [1]. In [20], interaction networks which accrue as aggregations of log files within the social tagging system BibSonomy are introduced and analyzed.

In contrast to previous work, the present work focuses on the question, whether a given social network gives rise to a notion of relatedness among its nodes and how different network variants, such as directedness and edge weights have an impact on the resulting network semantics. The proposed methodology is applied to different networks and structural similarity metrics, giving new insights into the semantics of those networks and their variants as well as the considered similarity metrics.

3 Network Data

Evidence Networks in BibSonomy BibSonomy is a social bookmarking system where users manage their bookmarks and publication references via *tag* annotations (i. e., freely chosen keywords). Most bookmarking systems incorporate additional relations on users such as "*my network*" in del.icio.us and "*friends*" in BibSonomy.

But beside those explicit relations among users, different relations are established implicitly by user interactions within the systems, e. g., when user u looks at user v's resources. As all of BibSonomy's log files were accessible, a broad range of interaction networks was available. In particular, we considered the directed *Friend-Graph*, containing an edge (u, v) iff user u has added user v as a friend, the directed *Copy-Graph* which contains an edge (u, v) with weight $c \in \mathbb{N}$, iff user u has copied c resources, i. e., a publication reference from user v and the directed *Visit-Graph*, containing an edge (u, v) with label $c \in \mathbb{N}$ iff user u has navigated c times to the user page of user v.

Evidence Networks in twitter We also considered the microblogging service twitter. Using twitter, each user publishes short text messages (called "*tweets*") which may contain freely chosen *hashtags*, i. e., distinguished words which are used for marking keywords or topics. Furthermore, users may "cite" each other by "retweeting": A user u retweets user v's content, if u publishes a text message containing "RT @v:" followed by (an excerpt of) v's corresponding tweet. Users may also explicitly follow other user's tweets by establishing a corresponding friendship-like link. For our analysis, we considered the directed *Follower-Graph*, containing an edge (u, v) iff user u follows the tweets of user v and the *ReTweet-Graph*, containing an edge (u, v) with label $c \in \mathbb{N}$ iff user u cited (or "retweeted") exactly c of user v's tweets.

Evidence Networks in flickr The flickr system focuses on organizing and sharing photographs collaboratively. Users mainly upload images and assign arbitrary tags but also interact, e. g., by establishing contacts or commenting on other users images. For our analysis we extracted the directed *Contact-Graph*, containing an edge (u, v) iff user u added user v to its personal contact list, the directed *Favorite-Graph*, containing an edge (u, v) with label $c \in \mathbb{N}$ iff user u added exactly c of v's images to its personal list of favorite images as well as the directed *Comment-Graph*, containing edge (u, v) with label $c \in \mathbb{N}$ iff user u posted exactly c comments on v's images.

General Structural Properties Table 1 summarizes major graph level statistics for the considered networks which range in size from thousands of edges (e. g., the Friend-Graph) to more than one hundred million edges (flickr's Contact-Graph). All networks obtained from BibSonomy are complete and therefore not biased by a previous crawling process. In return, effects induced by limited network sizes have to be considered.

Table 1. High level statistics for all networks with density d, the number of strongly connected components #scc and the size of the largest strongly connected component SCC.

	$\|V_i\|$	$\|E_i\|$	d	#scc	SCC
Copy	$1,427$	$4,144$	$2 \cdot 10^{-3}$	$1,108$	309
Visit	$3,381$	$8,214$	10^{-3}	$2,599$	717
Friend	700	$1,012$	$2 \cdot 10^{-3}$	515	17
ReTweet	$826,104$	$2,286,416$	$3,4 \cdot 10^{-6}$	$699,067$	$123,055$
Follower	$1,486,403$	$72,590,619$	$3,3 \cdot 10^{-5}$	$198,883$	$1,284,201$
Comment	$525,902$	$3,817,626$	$1,4 \cdot 10^{-5}$	$472,232$	$53,359$
Favorite	$1,381,812$	$20,206,779$	$1,1 \cdot 10^{-5}$	$1,305,350$	$76,423$
Contact	$5,542,705$	$119,061,843$	$3,9 \cdot 10^{-6}$	$4,820,219$	$722,327$

4 Analysis of Network Semantics

In Section 3, we introduced various explicit and implicit interaction networks from different applications. In this section, we tackle the problem of assessing the "meaning" of relations among pairs of vertices within such a network. This analysis then gives insights into the question, whether and to which extent the networks give rise to a common notion of *semantic relatedness* among the contained vertices. For this, we apply an experimental methodology, which was previously used for assessing semantical relationships within co-occurrence networks [21]. The basic idea is simple: We consider well founded notions of relatedness, which are naturally induced by external properties of the corresponding vertex sets, as, e. g., similarity of the applied tag assignments in BibSonomy or geographical distance between users in twitter. We than compute for each pair of vertices within a network these "semantic" similarity metrics and correlate them with different measures of structural similarity in the considered network.

4.1 Vertex Similarities

Below, we apply two well-established similarity functions in corresponding unweighted variants, namely the cosine similarity COS and the Jaccard Index JC as well as the corresponding weighted variants $\widetilde{COS}$ and $\widetilde{JC}$, following the presentation in [22].

Additionally we apply a modification of the *preferential PageRank* which we adopted from our previous work on folksonomies [5]: For a column stochastic adjacency matrix A and damping factor α, the *global* PageRank vector $\boldsymbol{w}$ with uniform *preference vector* $\boldsymbol{p}$ is given as the fixpoint of $\boldsymbol{w} = \alpha A\boldsymbol{w} + (1-\alpha)\boldsymbol{p}$. In case of the *preferential PageRank* for a given node i, only the corresponding component of the preference vector is set. For vertices x, y we set accordingly $PPR(x, y) := \boldsymbol{w}_{(x)}[y]$, that is, we compute the preferential PageRank vector $\boldsymbol{w}_{(x)}$ for node x and take its y'th component. We calculate the adopted preferential PageRank score by subtracting the global PageRank score PR from the preferential PageRank score in order to reduce frequency effects and set

$$PPR{+}(x, y) := PPR(x, y) - PR(x, y).$$

4.2 Semantic Reference Relations

For assessing the semantic similarity of two nodes within a network, we look for external properties which give rise to a well founded notion of relatedness. In the following, we consider the similarity of users based on the applied tags in BibSonomy and flickr, as well as the applied hashtags in twitter. We also consider geographical distance of users in twitter and flickr.

Tag Similarity In the context of social tagging systems like BibSonomy, the cosine similarity is often used for measuring semantic relatedness (see, e. g., [2]).

We compute the cosine similarity in the vector space $\mathbb{R}^T$, where, for user u, the entries of the vector $(u_1, \ldots, u_T) \in \mathbb{R}^T$ are defined by $u_t := w(u, t)$ for tags t where $w(u, t)$ is the number of times user u has used tag t to tag one of her resources (in case of BibSonomy and flickr) or the number of times user u has used hash tag t in one of her tweets (in case of twitter).

Geographical Distance In twitter and flickr, users may provide an arbitrary text for describing his or her location. Accordingly, these location strings may either denote a place by its geographic coordinates, a semi structured place name (e. g., "San Francisco, US"), a colloquial place name (e. g., "Motor City" for Detroit) or just a fantasy name. Also the inherent ambiguity of place names (consider, e. g., "Springfield, US") renders the task of *exactly* determining the place of a user impossible. Nevertheless, by applying best matching approaches, we assume that geographic locations can be determined up to a given uncertainty and that significant tendencies can be observed by averaging over many observations.

We used Yahoo!'s Placemaker™ API[3] for matching user provided location strings to geographic locations with automatic place disambiguation. In case of flickr, we obtained geographic locations for $320,849$ users and in case of twitter for $294,668$ users. Geographical distance of users is then simply given by the distance of the centroids for the correspondingly matched places.

Please note that geographic distance correlates with many secondary notions of relatedness between users, such as, e. g., language, cultural background and habits.

[3] http://developer.yahoo.com/geo/placemaker/

4.3 Grounding of Shortest Path Distance

For analyzing the interdependence of *semantic* and *structural* similarity between users, we firstly consider a very basic measure of structural relatedness between two nodes in a network, namely their respective shortest path distance. We ask, whether users which are direct neighbors in an evidence network tend to be more similar than distant users. That is, for every shortest path distance d and every pair of nodes u, v with a shortest path distance d, we calculated the average corresponding similarity scores $COS(u, v)$, $JC(u, v)$, $PPR(u, v)$ with variants and geographic distance. To rule out statistical effects, we repeated for each network G the same calculations on shuffled null model graphs.

Semantic Similarity Figure 1 shows the resulting plots for each considered network separately. Though the obtained average similarity scores vary greatly in magnitude for different networks (e. g., a maximum of 0.22 for the Friend-Graph in BibSonomy compared to a maximum of 0.1 for the Visit-Graph), they also share a common pattern: Direct neighbors are in average significantly more similar than distant pairs of users. And with a distance of two to three, users tend to be less similar than in average (in case of the ReTweet graph, users are more similar than in average up to a distance of eight). For the Visit-Graph, the Comment-Graph, the Follower-Graph and the ReTweet graph, the average similarity scores approach the global average similarity again. For distances around a network's diameter, the number of observations is too small, resulting in less pronounced tendencies for very distant nodes.

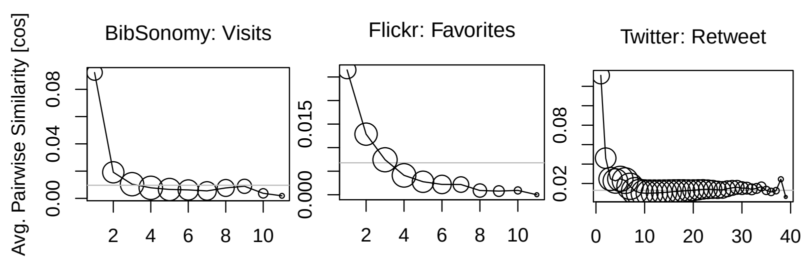

Fig. 1. Average pairwise cosine similarity based on the users' tag assignments relative to the shortest path distance in the respective networks where the global average is depicted in gray and the point size scales logarithmically with the number of pairs.

Geographic Distance For average geographic distances of users in flickr and twitter, we repeated the same calculations, as depicted in Figure 2. Firstly, we note the overall tendency, that direct neighbors tend to be located more closely than distant pairs of users within a network. Additionally, the average geographic distance of users then approaches the global average, and increases again after a certain plateau. As for the ReTweet-Graph, the average geographic distance remains at the global average level, once reached at a shortest path distance of ten.

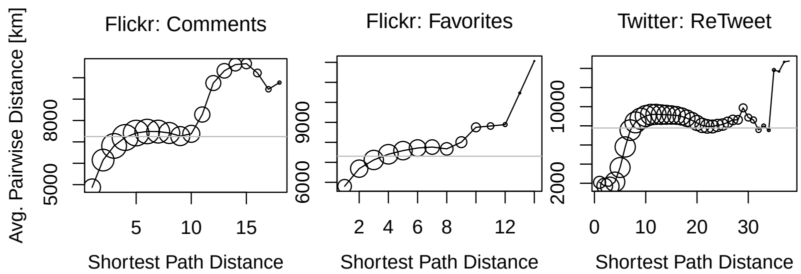

Fig. 2. Shortest path distance vs. average pairwise geographic distance in flickr. The global average is depicted in gray and the point size scales logarithmically with the number of pairs.

Discussion It is worth emphasizing, that in all considered evidence networks, the relative position of users already gives rise to a semantically grounded notion of relatedness, even in case of implicit networks, which are merely aggregated from usage logs as, e. g., the Visit-Graph. But one has to keep in mind that all observed tendencies are the result of averaging over a very large number of observations (e. g., $34, 282, 803, 978$ pairs of nodes at distance four in the Follower-Graph). Therefore, we cannot deduce geographic proximity from topological proximity for a given pair of users, as even direct neighbors in the Follower-Graph are in average located $4, 000$ kilometers apart from each other. But the proposed analysis aims at revealing semantic tendencies within a network and for comparing different networks (e. g., the Retweet-Graph better captures geographic proximity of direct neighbors in the graph). The experimental setup also allows to assess the impact of certain network variations, such as weighted and unweighted or directed and undirected networks, as exemplified in Section 4.5.

4.4 Grounding of Structural Similarity

We now turn our focus towards different measures of structural similarity for nodes within a given network. There is a broad literature on such similarity metrics for various applications, such as link prediction [15] and distributional semantics [6, 21]. We thus extend the question under consideration in Section 4.3, and ask, which measure of structural similarity best captures a given semantically grounded notion of relatedness among users. In the scope of the present work, we consider the cosine similarity and Jaccard index, which are based only on the direct neighborhood of a node as well as the (adjusted) preferential PageRank similarity which is based on the whole graph structure (refer to Section 4.1 for details).

Ultimately, we want to visualize correlations among structural similarity in a network and semantic similarity, based on external properties of nodes within it. We consider, again, semantical similarity based on users' tag assignments in BibSonomy, flickr and hash tag usage in twitter as well as geographic distance of users in flickr and twitter. In detail: For a given network $G = (V, E)$ and structural similarity metric S, we calculate for every pair of vertices $u, v \in V$ their structural similarity $S(u, v)$ in G as well as their semantic similarity and geographic distance. For visualizing correlations, we create plots with structural similarity at the x-axis and semantic similarity at the y-axis. As plotting the raw data points is computationally infeasible (in case of the Contact-Graph $30, 721, 580, 000, 000$ data points), we binned the x-axis and calculated average semantical similarity scores per bin. As the distribution of structural similarity scores is highly skewed towards lower similarity scores (most pairs of nodes have very low similarity scores), we applied logarithmic binning, that is, for a structural similarity score $x \in [0, 1]$ we determined the corresponding bin via $\lfloor \log(x \cdot b^N) \rfloor$ for given number of bins N and suitable base b. Pragmatically, we determined the base relative to the machine's floating point precision ϵ resulting in $b := \epsilon^{\frac{-1}{N}}$.

Semantic Similarity Figure 3 shows the obtained results for each considered network separately. We firstly note, that the cosine similarity metric and the Jaccard index are highly correlated. Secondly, the adjusted preferential PageRank similarity consistently outperforms the other similarity metrics with respect to magnitude and monotonicity (except for BibSonomy's Friend-Graph and flickr's Contact-Graph).

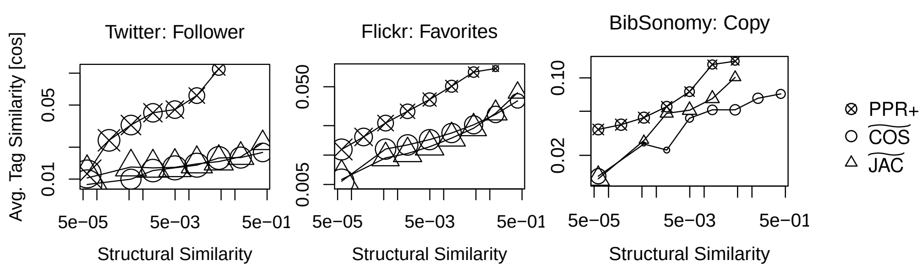

Fig. 3. Average pairwise semantic similarity based on tags users assigned to resources in BibSonomy and flickr or hash tag usage in twitter, relative to different structural similarity scores in the corresponding networks. The point size scales logarithmically with the number of pairs.

Geographic Distance As for geographic distances, Figure 4 shows the observed correlations for structural similarity in the different evidence networks and the corresponding average pairwise distance. In all but flickr's Favorite-Graph, for both local neighborhood based similarity metrics *COS* and JC, the average distance first decreases, but then increases again. This behavior is most pronounced in twitter's ReTweet-Graph. In the Favorite-Graph, both *COS* and JC monotonically decrease with increasing similarity score. On the other hand, the average distance decreases monotonically with increasing preferential PageRank score *PPR* consistently in all considered networks, except the ReTweet-Graph, where the average distance stays at a level of around 2.000 kilometers for similarity scores > 0. Generally (except for the ReTweet-Graph), it yields average distance values which are magnitudes below those obtained via the local similarity metrics.

Discussion Again, the obtained results only point at tendencies of the considered similarity metrics in capturing geographic proximity by means of structural similarity. Nevertheless, the adjusted preferential PageRank similarity consistently outperforms the other considered metrics. We therefore conclude that from all considered similarity metrics, the adjusted preferential PageRank similarity best captures the notion of

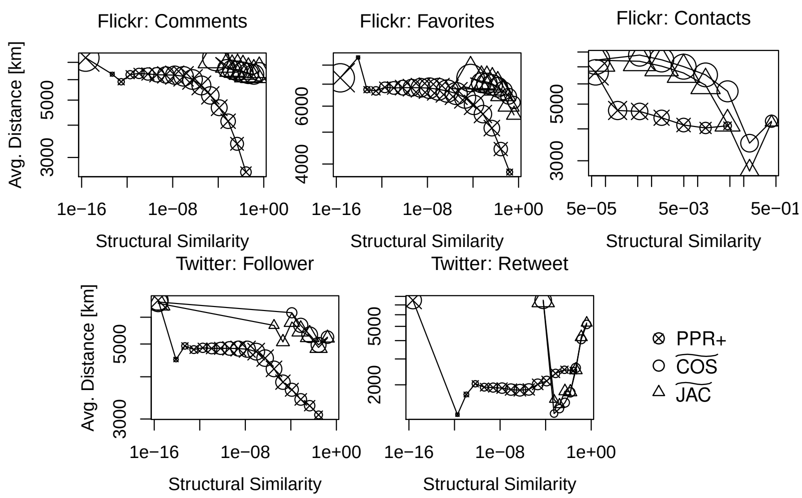

Fig. 4. Average pairwise distance relative to different structural similarity scores in the corresponding networks. The point size scales logarithmically with the number of pairs.

geographic proximity. This is especially of interest, as the geographic proximity is a prior for many properties users may have in common, such as, e. g., language, cultural background or habits. twitter's ReTweet-Graph seems to encompass the strongest geographic binding, as indicated in the relative low average distance for direct neighbors (cf. Figure 2 and the overall low average distance for higher preferential PageRank similarity scores (cf. Figure 4). Of course, other established similarity metrics (e. g., [6, 7, 12]) can be applied as well and are the subject of future considerations.

4.5 Case Study: BibSonomy

Most social networks are very sparse and in case of directed networks, dropping the direction of edges is a way of increasing a network's density. This might be of interest, e. g., for calculating a similarity function like COS, which is based on a node's neighborhood. The rational would be, that with a more dense adjacency matrix, more non-zero similarity scores are obtained.

The Impact of Directions We apply the semantic correlation analysis from Section 4.4 to assess the impact of dropping edge directions in the considered evidence networks in BibSonomy.

Figure 5 shows the corresponding plots, where the average similarity scores for the corresponding undirected networks are depicted in gray. The impact of dropping the directedness varies greatly among the different networks and similarity metrics. Firstly, the semantics of the cosine similarity in the Visit-Graph changes dramatically, by showing negative correlations with the (average) semantic similarity score in case of the undirected network. In the other networks, the cosine similarity's average semantic similarity scores are mainly reduced in magnitude.

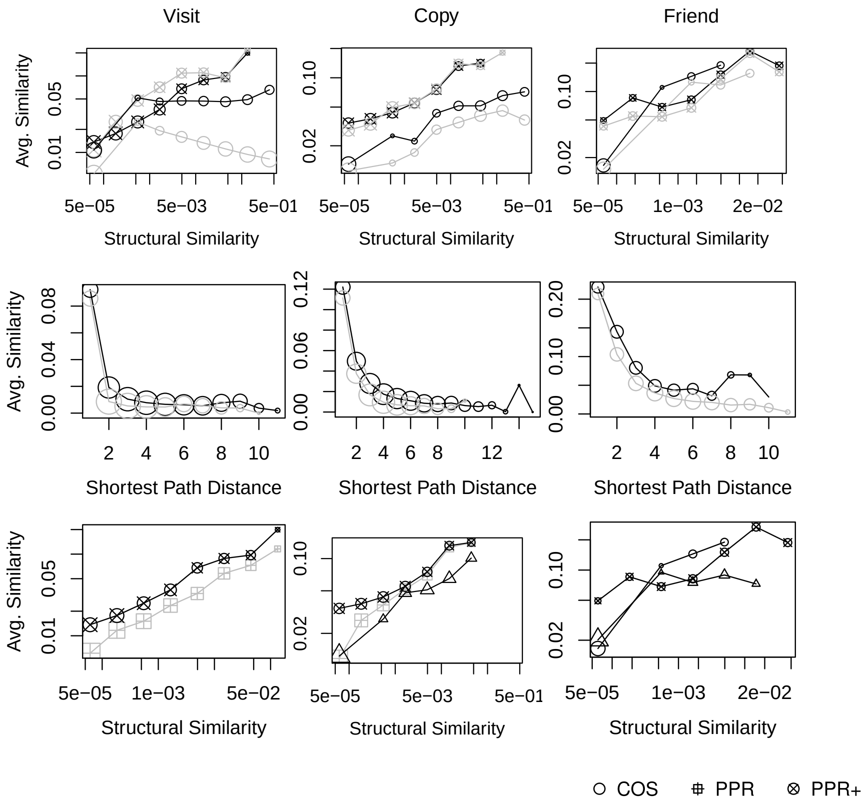

Fig. 5. Average pairwise semantic similarity in BibSonomy, relative to different structural similarity scores (upper row) and relative to the shortest path distance (mid row) in the corresponding directed and undirected networks. The impact considering the preference PageRank relative to the global PageRank is shown in the bottom row. Results for the undirected network variants are depicted in gray.

Considering the adjusted preferential PageRank similarity, no impact on the semantics can be observed in the Copy-Graph, a nearly constant decrease in the Friend-Graph, whereas in the Visit-Graph, the corresponding average semantic similarity is mostly increased, loosing monotonicity though.

In Figure 5, the average semantic similarity per shortest path distance in a network is also contrasted to the respective undirected variant. The undirected networks consistently show lowered average semantic similarity per shortest path distance.

Discussion With the preceding analysis, we exemplified, how the proposed experimental set up can be used for assessing the impact of changing certain network parameters, such as the directedness of edges. We conclude, that for the considered networks in BibSonomy the direction of edges significantly contributes to the network semantics and should not be dropped at all.

The Impact of Global PageRank In Section 4.1, we proposed to adjust the preferential PageRank similarity PPR by subtracting the global PageRank component wise. Using examples from the networks obtained from BibSonomy, we show, that this adjustment significantly increases the corresponding average semantic similarity in Figure 3.

5 Conclusion & Future Work

With the present work, we introduced an experimental framework for assessing the semantics of social networks. The proposed methodology has a broad range of applications, such as *user recommendation* or *community mining* tasks, as it allows semantically grounded pre processing of given networks (e. g., merging different small networks, scaling edge weights, selecting certain groups of users or directedness of networks). The conducted experiments give insights into the semantics of evidence networks from flickr, twitter and BibSonomy and well known similarity metrics. Additionally, the impact of directedness of a network and adjusting the preferential PageRank with the global PageRank is analyzed for the networks obtained from BibSonomy.

Ultimately, the proposed experimental setup allows to formulate the assessment of semantic user relatedness as a regression task, which will be subject to future work.

References

1. Brown, C., Nicosia, V., Scellato, S., Noulas, A., Mascolo, C.: Where Online Friends Meet: Social Communities in Location-based Networks. In: Proc. Sixth International AAAI Conference on Weblogs and Social Media (ICWSM 2012). Dublin, Ireland (2012)
2. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic grounding of tag relatedness in social bookmarking systems. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T.W., Thirunarayan, K. (eds.) The Semantic Web – ISWC 2008, Proc.Intl. Semantic Web Conference 2008. LNAI, vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
3. Cohen, T., Widdows, D.: Empirical distributional semantics: methods and biomedical applications. J Biomed Inform 42(2), 390–405 (Apr 2009)
4. Grefenstette, G.: Finding semantic similarity in raw text: The deese antonyms. In: Fall Symposium Series, Working Notes, Probabilistic Approaches to Natural Language. pp. 61–65 (1992)
5. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Sure, Y., Domingue, J. (eds.) The Semantic Web: Research and Applications. Lecture Notes in Computer Science, vol. 4011, pp. 411–426. Springer, Heidelberg (June 2006)
6. Islam, A., Inkpen, D.: Second order co-occurrence pmi for determining the semantic similarity of words. In: Proc. of the Int. Conference on Language Resources and Evaluation (LREC 2006). pp. 1033–1038 (2006)
7. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: Proc. of the eighth ACM SIGKDD int. conference on Knowledge discovery and data mining. pp. 538–543. KDD '02, ACM, New York, NY, USA (2002)
8. Kaltenbrunner, A., Scellato, S., Volkovich, Y., Laniado, D., Currie, D., Jutemar, E.J., Mascolo, C.: Far From the Eyes, Close on the Web: Impact of Geographic Distance on Online Social Interactions. In: Proc. ACM SIGCOMM Workshop on Online Social Networks (WOSN 2012). Helsinki, Finland (2012)

9. Kulshrestha, J., Kooti, F., Nikravesh, A., Gummadi, K.: Geographic dissection of the twitter network (2012)
10. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. pp. 591–600. ACM (2010)
11. Landauer, T., Dumais, S.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104(2), 211 (1997)
12. Leicht, E.A., Holme, P., Newman, M.E.J.: Vertex similarity in networks (2005), cite arxiv:physics/0510143
13. Lesk, M.: Word-word associations in document retrieval systems. American documentation 20(1), 27–38 (1969)
14. Li, N., Chen, G.: Analysis of a Location-Based Social Network. In: Proc. International Conference on Computational Science and Engineering. pp. 263–270. CSE '09, IEEE Computer Society, Washington, DC, USA (2009)
15. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. of the American society for inf. science and technology 58(7), 1019–1031 (2007)
16. Lü, L., Jin, C., Zhou, T.: Similarity index based on local paths for link prediction of complex networks. Physical Review E 80(4), 046122 (2009)
17. Lü, L., Zhou, T.: Link prediction in weighted networks: The role of weak ties. EPL (Europhysics Letters) 89, 18001 (2010)
18. McGee, J., Caverlee, J.A., Cheng, Z.: A Geographic Study of Tie Strength in Social Media. In: Proc. 20th ACM International Conference on Information and Knowledge Management. pp. 2333–2336. CIKM '11, ACM, New York, NY, USA (2011)
19. Mitzlaff, F., Atzmueller, M., Benz, D., Hotho, A., Stumme, G.: Community Assessment using Evidence Networks. In: Analysis of Social Media and Ubiquitous Data. LNAI, vol. 6904 (2011)
20. Mitzlaff, F., Benz, D., Stumme, G., Hotho, A.: Visit Me, Click Me, Be My Friend: An Analysis of Evidence Networks of User Relationships in Bibsonomy. In: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia. Toronto, Canada (2010)
21. Mitzlaff, F., Stumme, G.: Relatedness of given names. Human Journal 1(4), 205–217 (2012)
22. de Sá, H., Prudencio, R.: Supervised link prediction in weighted networks. In: Neural Networks (IJCNN), The 2011 Int. Joint Conference on. pp. 2281–2288. IEEE (2011)
23. Sadilek, A., Kautz, H., Bigham, J.P.: Finding Your Friends and Following Them to Where You Are. In: Proc. Fifth ACM International Conference on Web Search and Data Mining. pp. 723–732. WSDM '12, ACM, New York, NY, USA (2012)
24. Scellato, S., Noulas, A., Lambiotte, R., Mascolo, C.: Socio-spatial properties of online location-based social networks. Proceedings of ICWSM 11, 329–336 (2011)
25. Turney, P.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In: Proc. of the 12th European Conference on Machine Learning. pp. 491–502. Springer-Verlag (2001)
26. Weng, J., Lim, E., Jiang, J., He, Q.: Twitterrank: Finding Topic-Sensitive Influential Twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 261–270. ACM (2010)