Soc. Netw. Anal. Min. (2014) 4:216
DOI 10.1007/s13278-014-0216-2

ORIGINAL ARTICLE

The social distributional hypothesis: a pragmatic proxy for homophily in online social networks

Folke Mitzlaff · Martin Atzmueller · Andreas Hotho · Gerd Stumme

Received: 16 September 2013 / Revised: 11 May 2014 / Accepted: 2 July 2014 / Published online: 22 August 2014

Abstract Applications of the Social Web are ubiquitous and have become an integral part of everyday life: Users make friends, for example, with the help of online social networks, share thoughts via Twitter, or collaboratively write articles in Wikipedia. All such interactions leave digital traces; thus, users participate in the creation of heterogeneous, distributed, collaborative data collections. In linguistics, the *Distributional Hypothesis* states that words with similar distributional characteristics tend to be semantically related, i.e., words which occur in similar contexts are assumed to have a similar meaning. Considering users as (social) entities, their distributional characteristics can be observed by collecting interactions in social web applications. Accordingly, we state the *social distributional hypothesis*: we presume, that users with similar interaction characteristics tend to be related. We conduct a series of experiments on social interaction networks from Twitter, Flickr, and BibSonomy and investigate the relatedness concerning the interactions, their frequency, and the specific interaction characteristics. The results indicate interrelations between structurally similarity of interaction characteristics and semantic relatedness of users, supporting the social distributional hypothesis.

Keywords Social networks · Social interactions · Social media · Analysis · Distributional semantics

This article is part of the Topical Collection on Social Systems as Complex Networks.

F. Mitzlaff (✉) · M. Atzmueller · G. Stumme
Knowledge and Data Engineering Group, University of Kassel, Kassel, Germany
e-mail: mitzlaff@cs.uni-kassel.de

M. Atzmueller
e-mail: atzmueller@cs.uni-kassel.de

G. Stumme
e-mail: stumme@cs.uni-kassel.de

A. Hotho
Data Mining and Information Retrieval Group, University of Wuerzburg, Wuerzburg, Germany
e-mail: hotho@informatik.uni-wuerzburg.de

1 Introduction

The rapid development of the Internet and the growing availability of mobile web access has catalyzed the development and use of social web applications. Using such online social networks, people interact with each other and maintain relationship, e.g., by sending private messages and establishing friendship in facebook. The thereby induced networks of user relatedness are a valuable source of information for different applications, considering, e.g., the task of recommending new acquaintances (Chiluka et al. 2011; Dong et al. 2012) or finding groups of related users (Newman and Girvan 2004; Kashoob et al. 2010; Atzmueller and Mitzlaff 2011) for targeting a commercial campaign.

However, users also interact *implicitly* with each other, e.g., by adding an other user's photograph to the personal list of favorite photographs in Flickr, or by visiting an other user's collection of bibliographic references in BibSonomy. In the end, using any social web application, users leave digital traces within the involved databases and server log files. Then, this information can be aggregated to implicit networks of user relatedness. We motivate such networks as *evidence networks*, with a continuum from explicit to implicit evidence of user relatedness and according traces. The use of such emerging network of user

relatedness in applications, e.g., for finding groups of users, can be justified by assuming underlying homophilic processes (McPherson et al. 2001), i.e., by assuming that users tend to interact with similar users. Yet, the mere collection of interaction data does not allow for deriving such a causal interdependence.

In this work, we propose the *social distributional hypothesis*, a pragmatic proxy for homophily which only considers statistical correlation between interaction characteristics and similarity of users, referring to the distributional hypothesis in linguistics, which states that words with similar distributional characteristics (i.e., words which occur in similar contexts) tend to be similar semantically (Harris 1954). In our context, this means, that users with similar interaction characteristics in applications of the social web tend to be related. In contrast to linguistic entities, such as words (Harris 1954) or tags (Markines et al. 2009; Cattuto et al. 2008), which are associated with certain semantics (e.g., the thing denoted by a word), users lack such fixed connotations. We hypothesize that users are related by definition if they interact [which is in line with Luhmann's sociological systems theory, where social systems are considered as systems of communication (Luhmann 1993)]. This seems plausible for interaction networks of explicit user relations (e.g., friendship networks) but is less obvious for implicit interaction networks which are aggregated from server log files.

For underpinning and grounding this hypothesis, we follow a statistical approach by collecting covariates of users which we consider as indicators of user relatedness. We consider a broad range of possible user interactions in social web applications, i.e., Twitter, Flickr, and BibSonomy, and analyze correlations between the derived interaction characteristics and external metrics of user similarity.

Specifically, we consider *geographic proximity* and *similarity of the applied tag vocabulary* of different users. In particular, we then consider the following three research questions:

1. Are people who interact more similar than those who do not interact directly?
2. Do people who interact more frequently tend to be more similar?
3. Do people who share similar interaction characteristics tend to be more similar?

Accordingly, we conduct a series of experiments on social interaction networks (cf. Sect. 3.3), derived from the respective social web applications mentioned above. The results of these experiments support a positive answer to our first considered research question; we observe higher average similarity scores for directly interacting users in all considered evidence networks. Furthermore, concerning the second research question, we observe the tendency of higher similarity scores and lower geographic distances for increasing interaction frequencies; this also suggests a positive answer to this question. Finally, with respect to the third research question, overall more similar users can be observed for higher structural similarity scores with respect to the considered evidence networks. This supports a positive answer to this research question.

The remainder of the paper is structured as follows: Sect. 2 discusses related work. After that, Sect. 3 presents the necessary notions of the applied social network analysis methods. Furthermore, Sect. 4 describes the applied datasets. Next, the results of the experiments are described in Sect. 5. Finally, Sect. 6 concludes with a summary and interesting options for future research.

2 Related work

Within this work we present an analysis of relational data among users of social web applications. Thereby we consider correlations between structural relational properties and the similarity of semantically motivated covariates of users, also considering implicitly accruing relational data (such as profile views). Accordingly, related work can be categorized in (1) the analysis of social networks in the context of web applications, (2) the analysis of implicit social ties, and (3) the analysis of user covariates in such data sets as well as interrelations with social ties.

(1) The analysis of online social media, in this context especially the interrelations among the involved actors have attracted a lot of attention during the last decades. A thorough analysis of fundamental network properties and interaction patterns in Twitter can be found in Kwak et al. (2010), constituting a reference for large-scale network properties of the Twitter networks which are considered within this work too. In Mislove et al. (2007), the contact network in Flickr is analyzed and compared with networks derived from other popular online social networking sites. In contrast, this work considers several network structures derived from Flickr, whose network properties are thoroughly analyzed in our preceding work (Mitzlaff et al. 2011, 2013), which is focused on structural interrelations among the different networks. All networks derived from BibSonomy are introduced and analyzed in Mitzlaff et al. (2010).

(2) Crandall et al. (2008) discuss similarity and social influence in online communities, providing the general idea that friends interact similarly. Their

results indicate that there are feedback effects between similarity between actors and future interactions. Within this work, we focus on similarity of users based on their applied *tag vocabulary* as well as based on their *geographic proximity*.
Based on the analysis of social networks in different tagging systems, Schifanella et al. (2010) investigate the relationship of topological closeness (in terms of the length of shortest paths) with respect to the similarity of the applied tag vocabulary of the respective user pairs. This analysis was already applied to the considered networks derived from BibSonomy in Mitzlaff et al. (2010). We further adopt this analysis to our data sets and extend it to the analysis of geographic proximity. Additionally, we also consider (more generally) different measures of structural similarity in the social networks with respect to tag-based similarity and geographic proximity. More related work on correlation of the probability of social ties and the geographic distance of the corresponding users can be found in Scellato et al. (2011), McGee et al. (2011), and Kaltenbrunner et al. (2012).

3. Another aspect of our work is the analysis of *implicit link structures* which implicitly accrue in a running Web 2.0 system. Krause et al. (2008) analyzed term co-occurrence networks in logfiles of internet search systems. They showed that the exposed structure is similar to a folksonomy. Inspired by this result, we analyze implicit networks derived from BibSonomy in Mitzlaff et al. (2010) and compare properties of implicit networks with those from explicitly established network links in BibSonomy. This idea is further applied to Twitter and Flickr in Mitzlaff et al. (2013) where all implicit networks considered in this work are introduced.

In the context of link prediction, other implicit networks are considered. In Leroy et al. (2010), a feature-based approach using implicit information for inferring interaction networks is presented. Eagle et al. (2009) describe an approach for reconstructing friendship relations from secondary (mobile phone) data. They show, that friendship links can be inferred with a high probability. Our work applies lower level correlation analysis, aiming at providing a base for the use of implicitly accruing interaction data in running web applications.

Finally, it is worth noting that correlations among user interactions and user similarity is strongly related to the concept of homophily (McPherson et al. 2001). Our work focuses on results which imply no causality, as in typical data sets, no causal dependencies can be derived. The analysis of causal effects in online networking sites is recently presented in van de Rijt et al. (2014), where data is collected in accordingly designed social experiments.

3 Background

In the following, we briefly introduce basic notions, terms, and measures used in this paper: We summarize these notions and terms with respect to graphs, explicit and implicit relations, and similarity measures in graphs. Finally, we introduce the concept of evidence networks as the basis for our analyses. For more details, we refer to standard literature, e.g., Diestel (2006), Newman (2003), and Gaertler (2004).

3.1 Basic concepts and notation

For modeling (social) networks, we use concepts and notations from the study of graphs, i.e., *graph theory*. In the following, we refer to standard literature, e.g., Diestel (2006), for a detailed introduction and discussion of graph theory.

We denote a graph by $G = (V, E)$ where E is the edge set and V the vertex set. A binary relation on a set V is a *relation* R as a subset $R \subseteq V \times V$. A relation R is naturally mapped to a directed graph $G_R := (V, R)$. We say that a relation R among individuals U is *explicit*, if $(u, v) \in R$ only holds, when at least one of u, v *explicitly* established a connection to the other (e.g., user u added user v *deliberately* as a friend in an online social network). We call R *implicit*, if $(u, v) \in R$ can be *derived* from a set of other relations, e.g., it holds as a side effect of the actions taken by u and v in a social application. Explicit relations are thus given by explicit links, e.g., existing links between users.

Many observations of network properties can be explained just by the network's degree distribution (Kolaczyk 2009). It is therefore important to contrast the observed property to the according result obtained on a random graph as a *null model* which shares the same degree distribution. If a single network G is considered, a corresponding null model $\overline{G}$ can be obtained by randomly replacing edges $(u_1, v_1), (u_2, v_2) \in E$ with (u_1, v_2) and (u_2, v_1), ensuring that these edges were not present in G beforehand. This process is typically repeated a multiple of the graph edge set's cardinality (Maslov and Sneppen 2002). For contrasting comparative observations within pairs of networks (G_1, G_2), a null model $\underline{G_2}$ can be obtained by permuting the vertex positions within G_2 as described in Butts and Carley (2005).

3.2 Vertex similarities

Similarity scores for pairs of vertices based only on the surrounding network structure have a broad range of applications, especially for the link prediction task (Liben-Nowell and Kleinberg 2007). In the following, we present all considered similarity functions, following the presentation given in de Sá and Prudencio (2011) which builds on the extensions of standard similarity functions for weighted networks from Murata and Moriyasu (2007).

The *Jaccard coefficient* measures the fraction of common neighbors

$$\mathrm{JAC}(x,y) := \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|},$$

where $\Gamma(u)$ denotes the set of direct neighbors for node $u \in V$ of a graph $G = (V, E)$. The Jaccard coefficient is broadly applicable and commonly used for various data mining tasks. For weighted networks the Jaccard coefficient becomes

$$\widetilde{\mathrm{JAC}}(x,y) := \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(y,z)}{\sum_{a \in \Gamma(x)} w(a,x) + \sum_{b \in \Gamma(y)} w(b,y)}.$$

The *cosine similarity* measures the cosine of the angle between the corresponding rows of the adjacency matrix, which for an unweighted graph can be expressed as

$$\mathrm{COS}(x,y) := \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)|} \cdot \sqrt{|\Gamma(y)|}},$$

and for a weighted graph, the weighted cosine similarity $\widetilde{\mathrm{COS}}(x,y)$ between nodes x and y is given by

$$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) w(y,z)}{\sqrt{\sum_{a \in \Gamma(x)} w(x,a)^2} \cdot \sqrt{\sum_{b \in \Gamma(y)} w(y,b)^2}}.$$

The *preferential (or preference) PageRank* similarity is based on the well-known PageRank (Brin and Page 1998) algorithm; with $m \times m$ column stochastic adjacency matrix A, damping factor α, and uniform *preference vector* $\boldsymbol{p} := (1/m, \ldots, 1/m)$, the *global* PageRank vector $\boldsymbol{w} =: \mathrm{PR}$ is given as the fixpoint of the following equation:

$$\boldsymbol{w} = \alpha A \boldsymbol{w} + (1 - \alpha)\boldsymbol{p}$$

In case of the *preferential PageRank* for a given set of nodes $\mathcal{I}$, only the corresponding components of the preference vector are set and we set accordingly $\mathrm{PPR}(\mathcal{I})$ to the fixpoint of the above equation with

$$p_i := \begin{cases} \frac{1}{|\mathcal{I}|}, & \text{if i} \in \mathcal{I} \\ 0, & \text{otherwise.} \end{cases}$$

3.3 Evidence networks

Throughout this work, we assume an all-embracing underlying structure of relatedness among people, which we call the *social constellation*. This relatedness of people can neither be measured nor proven directly, but serves as a working hypothesis for justifying further assumptions. We consider digital traces of user interaction in social web applications as manifestation of the underlying social constellation and hence call aggregated networks of user relatedness "evidence networks", in reference to Hornby et al. (1974), where evidence is defined as "anything that gives reason for believing something, that makes clear or proves something". This twofold definition of the term "evidence" corresponds to the range of explanatory power of implicit user interactions, such as profile visits, in contrast to explicitly established friendship links in online social networks. Exploiting this kind of information is motivated by assuming certain underlying homophilic processes (McPherson et al. 2001), i.e., that users tend to interact with similar users. Interaction data can then provide indicators for statistical associations. Figure 1a shows a fictitious simplified social constellation for four given users of Twitter. *Bob* and *Ken* are friends, while *Bob* and *Larry* are brothers and, finally, *Ken* and *Eddie* are assumed being colleagues (we thus intentionally ignore further relations, such as *Larry* and *Eddie* preferring the color "green"). While Fig. 1b shows an evidence network derived from Twitter's Follower graph, Fig. 1c shows a different evidence network for the same set of users, which is derived from Twitter's ReTweet graph (cf. Sect. 4.1).

Formally, the social constellation can be modelled as a sequence $(R_1, R_2, \ldots)$ of relations $R_i \in \mathbb{R}^{n \times n}$, where n denotes the number of users. That is, there are countable many relations R_i defined on the set of users. For example, one relation $R_{\text{Parenthood}}$ may capture the concept of person u being parent of person v, while another relation R_{Work} may capture the notion of u and v being colleagues.

Considering some social application with a given social network N, we consider N as a sample of a subset

$$R_S = \bigcup_J \{R_j \mid j \in J\}, \quad \text{with } J \subseteq \mathbb{N},$$

that is, each observed edge (u, v) in N hints at u and v being related, without actually knowing the *nature* of this relationship. We call such a network an *evidence network* of user relatedness. We also call evidence networks which are based on user interactions (such as, e.g., sending messages) *social interaction network*.

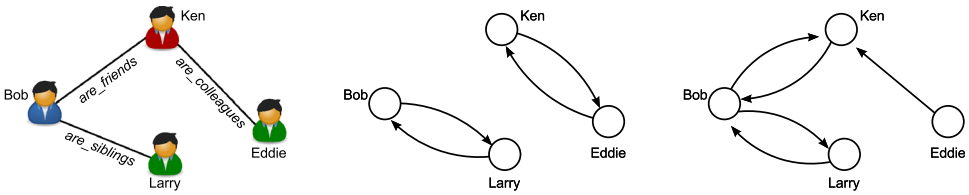

Fig. 1 Example for a given social constellation with corresponding evidence networks of user relatedness in Twitter, namely the Follower graph and the ReTweet graph

(a) Social Constellation **(b)** Follower graph **(c)** ReTweet graph

4 Data

For conducting our experiments, we aggregated various explicit and implicit evidence networks of user relatedness obtained from different applications from the social web and collected external properties of according user nodes from which we derive measures of user similarity. Subsequently, we first describe the considered evidence networks and summarize corresponding general high-level statistics (these networks are thoroughly analyzed and compared in Mitzlaff et al. (2013)). After that, we present the collected data sets of external user properties.

4.1 Network data

4.1.1 Networks in Twitter

Firstly, we consider 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 following networks:

- The *Follower graph* is an explicit evidence network, given by a directed graph containing an edge (u, v) iff user u follows the tweets of user v.
- The *ReTweet graph* is an implicit evidence network, given by a directed graph; it contains an edge (u, v) with weight $c \in \mathbb{N}$ iff user u "retweeted" exactly c of user v's tweets.

We extracted Twitter 's ReTweet graph from a Twitter data set, published in Yang and Leskovec (2011), which is estimated to cover 20–30 % of all public tweets published on Twitter during 2009-06-01 to 2009-12-31. Additionally, we used the follower network as made available in Maslov and Sneppen (2002) which was crawled during the time period 2009-06-01 until 2009-09-24, containing more than 1.4 billion following relations. For our analysis, we only considered users which were also present in the tweets data set.

4.1.2 Networks in Flickr

Flickr 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 images of other users. For our analysis, we extracted the following networks:

- The *Contact graph* is an explicit evidence network given by a directed graph; it contains an edge (u, v) iff user u added user v to its personal contact list.
- The *Favorite graph* is an implicit evidence network given by a directed graph containing an edge (u, v) with weight $n \in \mathbb{N}$ iff user u added exactly n of v's images to its personal list of favorite images.
- The *Comment graph* is an implicit evidence network; the directed graph contains an edge (u, v) with a weight $c \in \mathbb{N}$ iff user u posted exactly c comments on v's images.

The Flickr networks were extracted from an own breadth-first crawl, which was conducted in April until June 2011. The search was regularly reseeded by randomly selecting a search term from a library catalogue search term data set[1] which was then used for querying images using Flickr's API.[2] In parallel all incident comments, users, contacts, and favorites were crawled.

Beside the aforementioned evidence networks, the considered Flickr data set consisted of $588,634$ photos for a set of $69,104$ users who applied $564,251$ different tags in $5,911,127$ tag assignments. Data sets obtained by breadth-first crawl techniques are known to be biased toward high-degree nodes (Gjoka et al. 2011) and likely underestimate link symmetry (Becchetti et al. 2006). This work aims at comparing structural characteristics of different networks within a given social constellation (e.g., on the set of users in Flickr) rather than characterizing the networks. However, the different networks obtained in Flickr were

[1] http://data.gov.au/1277.

[2] http://www.flickr.com/services/api/.

crawled in parallel. Thus, induced biases have a comparable impact on all considered networks.

4.1.3 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[3] and "*friends*" in BibSonomy[4]. Each such network is connected with a certain functionality, e.g., for restricting access to certain resources or for allowing messages to be sent. Nevertheless, during the period of the evaluation, $<5\,\%$ of BibSonomy's friendship links were used according to its functional intention. All remaining links were used for expressing some sort of affiliation to other users and we accordingly expect that those networks also have a certain "social meaning".

- The *Friend graph* is a directed graph containing an edge (u, v) iff user u has added user v as a friend. In BibSonomy, the only purpose of the friend graph so far is to restrict access to selected posts so that only users classified as "friends" can observe them.
- The *Group graph* is an undirected graph containing an edge $\{u, v\}$ iff user u and v share a common group, e.g., defined by a certain research group or a special interest group.

Due to its limited size, we excluded the network obtained from BibSonomy's follower feature which enables users to monitor new posts of other users.

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. Using the BibSonomy's log files, a broad range of interaction networks were available.

- The *Click graph* is a directed graph containing an edge (u, v) iff user u has clicked on a link on the user page of user v.
- The *Copy graph* is a directed graph containing an edge (u, v) iff user u has copied a resource, i.e., a publication reference from user v.
- The *Visit graph* is a directed graph containing an edge (u, v) iff user u has navigated to the user page of user v.

Each implicit graph is given a weighting function counting certain events (e.g., the number of posts which user u has copied from v in case of the Copy graph). Our primary resource is an anonymized dump of all public bookmark and publication posts until January 25, 2010. It consists of 175,521 tags, 5,579 users, 467,291 resources, and 2,120,322 tag assignments. The dump also contains friendship relations modeled in BibSonomy among 700 users. Furthermore, we utilized the "*click log*" of BibSonomy, consisting of entries which are generated whenever a logged-in user clicked on a link in BibSonomy. A log entry contains the URL of the currently visited page together with the corresponding link target, the date and the user name.[5] For our experiments, we considered all click log entries until January 25, 2010. Starting in October 9, 2008, this dataset consists of 1,788,867 click events in total. We finally considered the corresponding apache web server log files, containing around 16 GB compressed log entries.

4.1.4 General structural properties

Table 1 summarizes major graph level statistics for the considered networks, which range in size from hundreds of edges (e.g., BibSonomy's 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, but effects induced by limited network sizes have to be considered.

Table 2 also shows the diameter, average path length, and the transitivity (also called clustering coefficient) for all considered networks. Except for the Group graph, the Friend graph and the ReTweet graph, all networks exhibit a comparable magnitude of these indices. While the Group graph and the Friend graph are characterized by a large transitivity, the ReTweet graph shows an unexpected high diameter and average path length.

Figure 2 breaks down the average to the distribution of path lengths. The Click graph and the Visit graph, for example, show a clear common distributional pattern as do the Copy graph, the Retweet graph, the Follower graph and the Favorite graph where both groups have a single cluster point around the graph's average path length.

4.2 Semantic reference relations

For assessing the semantic similarity of two users within a network, we look for external properties which give raise 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 (cf. Sect. 4.1). We also consider geographical distance of users in Twitter and Flickr.

[3] http://delicious.com/network/.

[4] http://www.bibsonomy.org/friends.

[5] Note: For privacy reasons a user may deactivate this feature.

Table 1 High-level statistics for all considered evidence networks, with the number of strongly connected components #SCC, the size of the largest strongly connected component |SCC| and the size of the largest weekly connected component |WCC|

	$\|V\|$	$\|E\|$	Density	#SCC	|SCC|	|WCC|
BibSonomy						
Copy	1,427	4,144	2×10^{-3}	1,108	309	1,339
Click	1,151	1,718	10^{-3}	963	150	1,022
Visit	3,381	8,214	10^{-3}	2,599	717	3,359
Group	550	6,693	2.2×10^{-3}	–	–	228
Friend	700	1,012	2×10^{-3}	515	17	238
Twitter						
ReTweet	826,104	2,286,416	3.4×10^{-6}	699,067	123,055	702,809
Follower	1,486,403	72,590,619	3.3×10^{-5}	198,883	1,284,201	1,485,356
Flickr						
Comment	525,902	3,817,626	1.4×10^{-5}	472,232	53,359	522,212
Favorite	1,381,812	20,206,779	1.1×10^{-5}	1,305,350	76,423	1,380,906
Contact	5,542,705	119,061,843	3.9×10^{-6}	4,820,219	722,327	5,542,703

Table 2 Path statistics with average path length (APL) for all networks where the Krackhardt Hierarchy (KH) values marked with an asterisk are estimated by repeatedly averaging over random samples of pairs of vertices

	Diameter	APL	Transitivity	Symm. links	KH
BibSonomy					
Copy	15	4.3	0.10	0.09	0.80
Click	15	4.8	0.02	0.12	0.88
Visit	11	3.9	0.01	0.12	0.81
Group	7	2.9	0.85	–	–
Friend	10	3.4	0.28	0.12	0.81
Twitter					
ReTweet	39	9.7	0.06	0.12	0.81*
Follower	13	3.3	0.01	0.55	0.12*
Flickr					
Comment	18	4.4	0.03	0.08	0.91*
Favorite	11	3.3	0.02	0.03	0.96*
Contact	8	2.9	0.05	0.46	0.87*

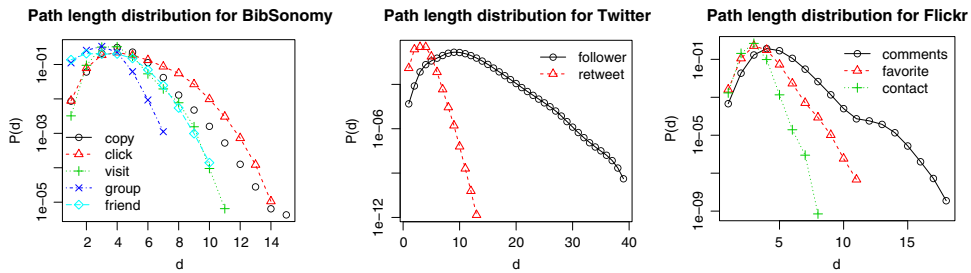

Fig. 2 Distribution of the shortest path lengths in the evidence networks with logarithmically scaled counts on the Y-axis

4.2.1 Tag-based similarity

In the context of social tagging systems like BibSonomy, the cosine similarity is often used for measuring semantic relatedness (Markines et al. 2009; Cattuto et al. 2008). We compute the cosine similarity in the vector space $\mathbb{R}^T$, where for user u, the entries of the vector $\boldsymbol{u} := (u_1, \ldots, u_T) \in \mathbb{R}^T$ 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). Each vector can be interpreted as a "semantic profile" of the underlying user, represented by the distribution of her tag usage. We then adopt the standard approach of information retrieval and compute in this vector space the cosine similarity between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ (cf. Sect. 3.2). This measure is thus independent of the length of the vectors. Its value ranges from -1 (for totally orthogonal vectors) to 1 (for vectors pointing into the same direction). In our case, the similarity

values lie between 0 and 1 because the vectors only contain positive numbers (Markines et al. 2009).

4.2.2 Geographical distance

In Twitter and Flickr, users may provide an arbitrary text for describing the user's home 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[6] for matching user provided location strings to geographic locations with automatic place disambiguation. In the case of Flickr, we obtained geographic locations for $320,849$ users and in case of Twitter for $294,668$ users. The 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.

5 Experiments

Within this section, we present the results of the experiments with respect to our three research questions, cf. Sect. 1 as applied on the different networks and semantically grounded utilizing the user similarity metrics described in Sect. 4.

Corresponding to our three research questions, we firstly consider the interdependence between interaction proximity and user similarity, secondly the impact of interaction frequencies and finally correlations between distributional interaction characteristics of users and according user similarity.

Please note that the standard error of mean is depicted in the diagram whenever appropriate by an according error bar. But in many cases, the number of corresponding observations is very high due to the nature of pairwise computations in network data (e.g., $30,721,580,000,000$ user pairs in case of Flickr's contact graph) and therefore the standard error often diminishes due to its normalization with the square root of the number of observations.

[6] http://developer.yahoo.com/geo/placemaker/ (November 2011).

5.1 Grounding of interaction proximity

Corresponding to the *first research question*, we consider the average pairwise covariate similarity of users (e.g., the average geographic distance) relative to the shortest path distance of the according user nodes within the network. 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 $\mathrm{COS}(u,v)$, $\mathrm{JAC}(u,v)$, $\mathrm{PPR}(u,v)$ with variants (cf. Sect. 3.2) and geographic distance. To rule out statistical effects, we repeated for each network G the same calculations on five independently generated corresponding null model graphs $\underline{G}$ (cf. Sect. 3.1) and depict the corresponding average results in gray. The analysis of average pairwise similarity scores relative to respective shortest path distances within a given network is based on Schifanella et al. (2010).

Then, we investigate, whether a negative correlation between the *average pairwise semantic similarity* and the corresponding *shortest path distance* of users in evidence networks can be observed.

5.1.1 Tag-based similarity

Figure 3 shows the resulting plots for each considered network based and BibSonomy as well as Flickr and Twitter, respectively. 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. Then, with a distance of two to four, users tend to be less similar than the average similarity score for all strongly connected pairs of nodes (which is depicted by a gray dashed line). In case of the ReTweet graph, users are more similar than in average up to a distance of eight. For distances around a network's diameter, the number of observations is very small, resulting in less pronounced tendencies for very distant vertex pairs. In all cases, the null model networks do not show an according interdependence between the shortest path distance and average user similarity, which for all distances approximates the global average.

5.1.2 Geographic distance

For average geographic distances of users in Flickr and Twitter, we repeated the same calculations, and show the obtained results in Fig. 4. We note the overall tendency that direct neighbors tend to be located more closely than distant pairs of users within a network. For all but the Follower graph and the ReTweet graph, the average

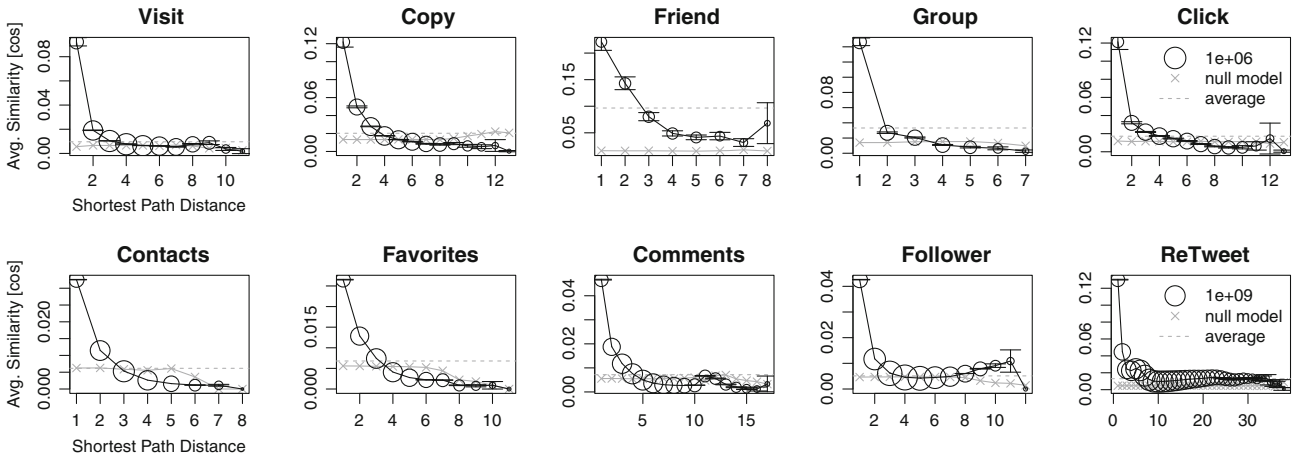

Fig. 3 Average pairwise cosine similarity based on the users' tag assignments relative to the shortest path distance in the respective networks, contrasted to corresponding results on the respective null model graphs $\underline{G}$ as depicted in *gray*

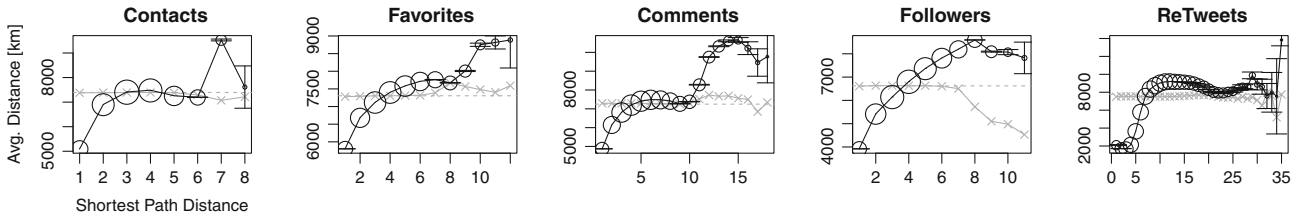

Fig. 4 Shortest path distance versus average pairwise geographic distance in Flickr. To rule out statistical effects, the geolocations of users where shuffled, as plotted in *gray*

geographic distance of users then approaches the global average for strongly connected node pairs, but after a certain plateau, increases again. In the Follower graph, the average geographic distance increases monotonically up to a shortest path distance of eight, remaining at the same average distance for higher distances (up to variance due to reduced number of observations). As for the ReTweet graph, the average geographic distance remains at the global average level, once reached at a shortest path distance of ten. Again, in the null model graphs, the average geographical distance approximates the global average for all shortest path distances, exhibiting no interdependence between distance in the interaction network and geographical distance.

5.1.3 Discussion

The results presented above support a positive answer to our first considered research question, by showing higher average similarity scores for directly interacting users in all considered evidence networks. It is worth emphasizing that the relative position of the users gives raise 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 km 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.

5.2 Grounding of interaction frequencies

With our *second research question* we want to investigate, whether the interaction frequency of user pairs correlates with semantic similarity of the incident users. For this, we show the *average semantic pairwise similarity* of users relative to the according interaction frequency.

Thus, we count the number of interactions per user pair and label the according edge in the corresponding evidence network accordingly. For example, in Twitter, we count for a pair of users (u, v), how often user u retweeted tweets of user v or in BibSonomy, how often user u has accessed user v's profile page. In the first case, one clearly expects that an increasing number of retweets increases the according hash tag similarity, as with each retweet, user u adopts part of user v's hash tags. In the latter case, this tendency is not that obvious, as a high profile page access frequency from u to v's profile page may just be the statistical result of v's high activity level in BibSonomy. To analyze and compare the impact of interaction frequencies within the different interaction networks, we consider the average semantic similarity with respect to the corresponding edge weight for each considered weighted network separately (accordingly, the explicit networks are not included in this experiment).

To account for the long-tailed distribution of edge weights and accordingly sparsely scattered observations for higher interaction frequencies, we applied logarithmic binning for calculating average semantic similarity scores. 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 a selected value of maximum precision $\epsilon := 10^{-8}$, resulting in $b := \epsilon^{\frac{-1}{N}}$. In the following, we present the obtained results first for the tag-based similarity in Twitter, Flickr and BibSonomy and then the geographical distance-based similarity for Twitter and Flickr.

5.2.1 Tag-based similarity

Figure 5 shows the average pairwise cosine similarity between the corresponding users' tag or hash tag context vectors for BibSonomy, Flickr and Twitter. As expected, for the Copy graph and the ReTweet graph, the correlation of interaction frequency and average pairwise similarity is most pronounced, as with copying a post in BibSonomy or retweeting a tweet in Twitter, most likely part of the originating tags are reused. But also the Visit, Click, and Favorite graph give rise to increasing average similarity scores with increasing number of interactions. For the Comment graph, the average similarity scores firstly show increasing, but then (starting at around 1,000 commented photographs) decreasing tendencies with respect to higher interaction frequencies. We assume that part of this pattern can be explained with artifacts due comment spam, e.g., by automatically generated entries, as the very high number of commented photographs (more than 5,000) per user pair suggests.

5.2.2 Geographic distance

For the average geographic distance, most notably, the ReTweet frequency shows strong geographical binding. Already for up to two retweets, the average geographical

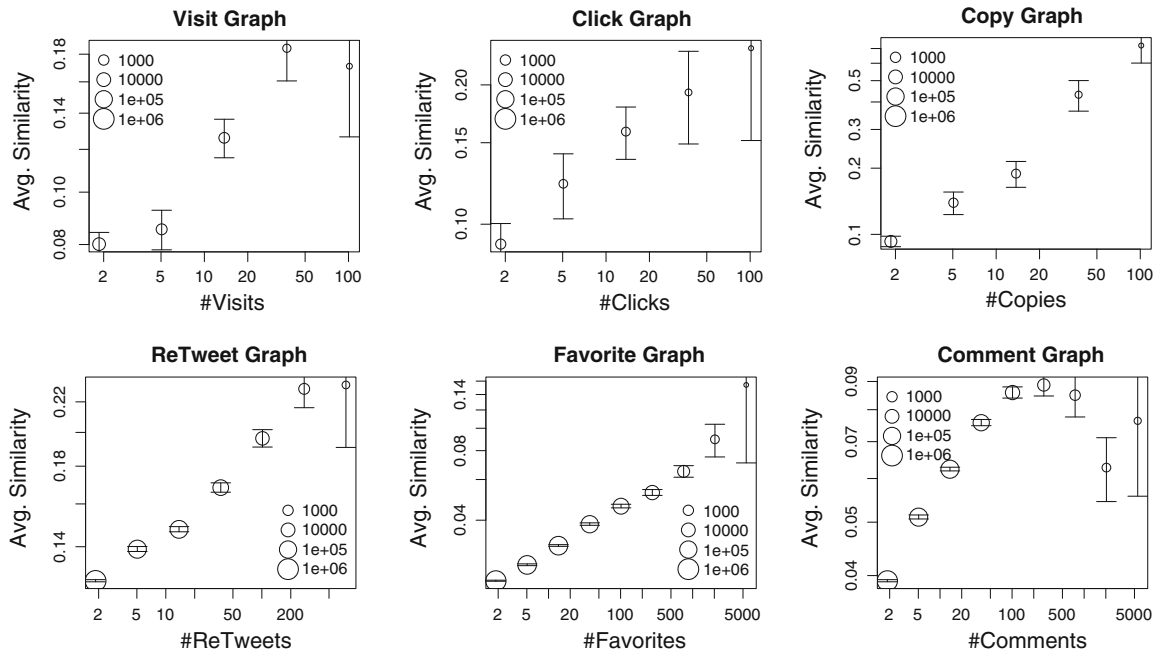

Fig. 5 Average pairwise cosine similarity of the respective tag context vectors for interacting users, relative to the corresponding interaction frequencies. Except for the Comment graph, users with higher interaction counts tend be more similar than users with low number interaction

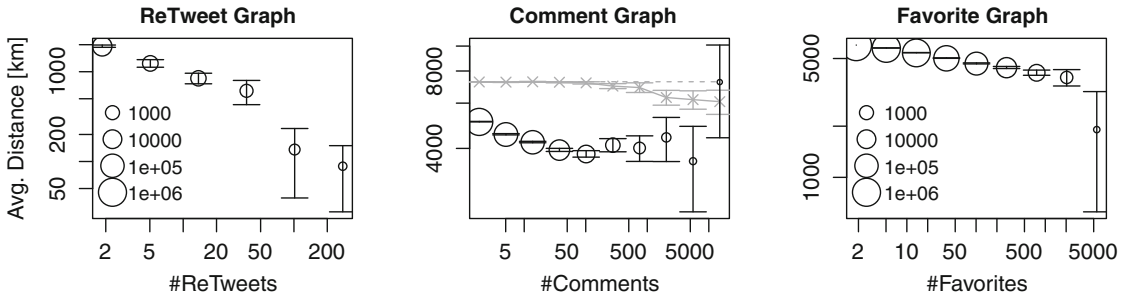

Fig. 6 Average pairwise geographic distance for interacting users, relative to the corresponding interaction frequencies. Except for the Comment graph, users with higher interaction counts tend be located more closely

distance drops below 2,000 km, in contrast to the global average pairwise distance of 7,400 km. For higher retweet counts, the average pairwise distance even drops below 200 kilometers. For the Favorite graph, the average pairwise geographical distance likewise tends to decrease for higher counts of favorite photographs, less pronounced than for the ReTweet graph though. Finally, the Comment graph exhibits the same pattern of dependency between interaction frequency and semantic similarity as for the tag-based similarity, by firstly showing a clear decreasing tendency for the average pairwise geographical distance which changes to an increasing tendency for higher retweet counts. Again, we attribute the latter increasing tendency to artifacts induced by automatic commenting processes (Fig. 6).

5.2.3 Discussion

Altogether, the observed tendencies of higher similarity scores and lower geographic distances for increasing interaction frequencies give evidence for a positive answer to the considered research question. Nevertheless, comparing the results obtained from the considered interaction networks, we note a significant difference in shape and magnitude of the respective average similarity curves.

The strongest relationship between interaction frequency and user similarity is observed in the ReTweet graph, both for the tag-based similarity (strongly biased by the retweeting process) and the geographical distance. While the former could be explained merely as an artifact induced by copying the retweeted message's hash tags, the latter shows that retweeting user pairs tend to be located more closely. 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.

But also the very implicit interaction of visiting a user's profile page in BibSonomy already gives rise to tendencies of higher user interrelationship for more intensively interacting users.

5.3 Grounding of structural interaction similarity

So far, we only considered basic structurally induced relations among nodes within a network, namely the interaction frequency with neighbors as well as the shortest path distance between pairs of nodes. Our *third research question* turns our focus toward further distributional measures of structural similarity for nodes within a given network, by analyzing correlations between such similarity metrics and measures of semantic similarity of users.

To address the third research question, we consider correlations between structural pairwise similarity of users within an evidence network and the corresponding pairwise covariate similarity. There is a broad literature on according similarity metrics for various applications, such as link prediction (Liben-Nowell and Kleinberg 2007) and distributional semantics (Islam and Inkpen 2006; Markines et al. 2009). We thus extend the question under consideration 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 both are based only on the direct neighborhood of a node, as well as the (differential) preference PageRank similarity which is based on the whole graph structure (cf. Sect. 3.2). Ultimately, we want to visualize correlations between structural similarity in a network and semantic similarity, based on external properties of nodes within it. Again, we consider 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.

We firstly consider tag-based similarity of users in BibSonomy, Flickr and Twitter and then geographical

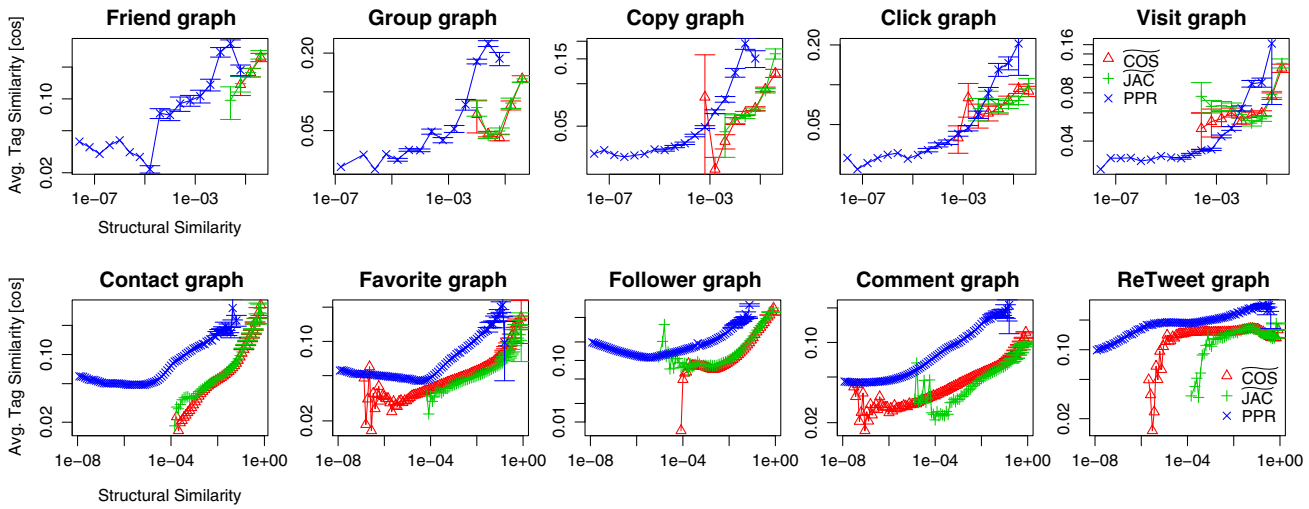

Fig. 7 Average pairwise semantic similarity based on tags which users assigned to resources in BibSonomy, relative to structural similarity scores in the corresponding evidence networks

distance of users in Flickr and Twitter. 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 toward lower similarity scores (most pairs of nodes have very low similarity scores), we applied logarithmic binning (cf. Sect. 5.2).

5.3.1 Semantic similarity

Figure 7 shows the obtained results for each considered network separately. We firstly note that the cosine similarity metric and the Jaccard index are highly correlated, whereby the Jaccard index shows slightly higher average semantic similarity scores for structurally more similar users than the cosine similarity in case of Flickr 's Contact graph and BibSonomy 's Copy graph. Secondly, the preferential PageRank similarity shows higher semantic similarity scores for all but the explicit Contact and Follower networks. For the Favorite and Follower graph, the preferential PageRank similarity indicates slightly negative correlation with the semantic similarity of users for lower structural similarity scores, but positive correlations for similarity scores $\geq 10^{-4}$.

5.3.2 Geographic distance

As for geographic distances, Fig. 8 shows the observed correlations for structural similarity in the different evidence networks and the corresponding average pairwise distance. In all but the Favorite and ReTweet graph, both local neighborhood-based similarity metrics COS and JAC, the average distance first decreases, but then increases again with higher similarity scores. In contrast, to Twitter's ReTweet graph capture, where both similarity metrics capture increasing geographic distance for more similar users. In the Favorite graph, both COS and JAC 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. In all but the Contact graph and the ReTweet graph, the preferential PageRank score indicates the lowest average distances for high similarity scores. As for the ReTweet graph, the preferential PageRank scores yield decreasing geographic distance at first (for scores in $[0, 10^{-5})$), but then increasing distances for higher similarity scores.

5.3.3 Discussion

Although different networks and similarity measures show deviating results in some cases, altogether more similar users can be observed for higher structural similarity scores in the considered evidence networks, giving support for a positive answer to the considered research question.

The obtained results thus point at tendencies of the considered similarity metrics in capturing tag-based semantics similarity and geographic proximity of users by means of structural similarity. Please note that the obtained average of the semantic similarity scores for higher structural similarity scores in the evidence network is significantly higher than the observed average semantic similarity score of directly interacting users (cf. Sect. 5.1) which indicates that structurally similar users are candidates for recommending new links within an application. Notably,

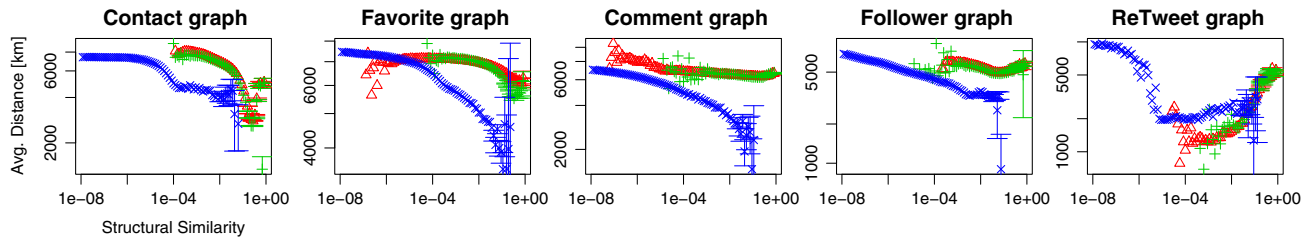

Fig. 8 Average pairwise distance relative to different structural similarity scores in the corresponding networks

the preferential PageRank similarity best captured in most cases, both tag bases similarity and geographic proximity of users.

6 Conclusions

Within this work, we presented the social distributional hypothesis, stating that users with similar interaction characteristics tend to be semantically related. For grounding this hypothesis, we considered three research questions, each of which pointing at different aspects of structurally induced notions of user relatedness in social interaction networks. These research questions were experimentally investigated for different traces of user interaction in social web applications, ranging from implicit profile page visits in BibSonomy to explicit Contact links in Flickr. These traces were used to build corresponding evidence networks of user relatedness. The conducted experiments affirm tendencies of interrelations between structurally similarity of interaction characteristics and semantic relatedness of users, supporting the social distributional hypothesis and thus justifying the use of even implicitly accruing social interaction networks for the analysis of user relatedness or for assessing the quality of user recommendation and community mining models.

Acknowledgments This work has been partially supported by the Commune project funded by the Hertie foundation.

References

Atzmueller M, Mitzlaff F (2011) Efficient descriptive community mining. In: Proceedings 24th international FLAIRS conference, AAAI Press, pp 459–464

Becchetti L, Castillo C, Donato D, Fazzone A, Rome I. (2006) A comparison of sampling techniques for web graph characterization. In: Proceedings of the workshop on link analysis (LinkKDD'06), Philadelphia, PA

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1):107–117

Butts CT, Carley KM (2005) Some simple algorithms for structural comparison. Comput Math Org Theory 11:291–305. doi:10.1007/s10588-005-5586-6.

Cattuto C, Benz D, Hotho A, Stumme G (2008) Semantic grounding of tag relatedness in social bookmarking systems. In: The Semantic Web—ISWC 2008, Proceedings of international semantic web conference 2008, LNAI, vol 5318. Springer, Heidelberg, pp 615–631

Chiluka N, Andrade N, Pouwelse J (2011) A link prediction approach to recommendations in large-scale user-generated content systems. In: Clough P, Foley C, Gurrin C, Jones G, Kraaij W, Lee H, Mudoch V (eds) Advances in information retrieval. Lecture notes in computer science, vol 6611. Springer, Berlin Heidelberg, pp 189–200

Crandall DJ, Cosley D, Huttenlocher DP, Kleinberg JM, Suri S (2008) Feedback effects between similarity and social influence in online communities. In: Proceedings of 14th ACM SIGKDD internatinal conference on knowledge discovery and data mining, ACM, pp 160–168

de Sá H, Prudencio R (2011) Supervised link prediction in weighted networks. In: The 2011 international joint conference on neural networks (IJCNN), pp 2281–2288. IEEE

Diestel R (2006) Graph theory. Springer, Berlin

Dong Y, Tang J, Wu S, Tian J, Chawla NV, Rao J, Cao H (2012) Link prediction and recommendation across heterogeneous social networks. In: Proceedings of the 2012 IEEE 12th international conference on data mining, ICDM'12. IEEE computer society, Washington, DC, USA, pp 181–190

Eagle N, Pentland A, Lazer D (2009) Inferring friendship network structure by using mobile phone data. Proc Natl Acad Sci 106(36):15274–15278. doi:10.1073/pnas.0900282106

Gaertler M (2004) Clustering. In: Brandes U, Erlebach T (eds) Network analysis, LNCS, vol 3418. Springer, Berlin, pp 178–215

Gjoka M, Kurant M, Butts CT, Markopoulou A (2011) Practical recommendations on crawling online social networks. IEEE J Select Areas Commun 29(9):1872–1892

Harris ZS (1954) Distributional structure. Word

Hornby AS, Cowie AP, Gimson AC, Lewis JW (1974) Oxford advanced learner's dictionary of current English, vol 1428. Cambridge Univ Press, Cambridge

Islam A, Inkpen D (2006) Second order co-occurrence PMI for determining the semantic similarity of words. In: Proceedings of the international conference on language resources and evaluation (LREC 2006), pp 1033–1038

Kaltenbrunner A, Scellato S, Volkovich Y, Laniado D, Currie D, Jutemar EJ, Mascolo C (2012) Far from the eyes, close on the web: impact of geographic distance on online social interactions. In: Proceedings ACM SIGCOMM workshop on online social networks (WOSN 2012) Helsinki, Finland

Kashoob S, Caverlee J, Kamath K (2010) Community-based ranking of the social web. In: Proceedings of the 21st ACM conference on hypertext and hypermedia

Kolaczyk E (2009) Statistical analysis of network data: methods and models. Springer Series in Statistics, p 386

Krause B, Jäschke R, Hotho A, Stumme G (2008) Logsonomy-social information retrieval with logdata. In: Proceedings 19th conference on hypertext and hypermedia, ACM, pp 157–166

Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web. ACM, pp 591–600

Leroy V, Cambazoglu BB, Bonchi F (2010) Cold start link prediction. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'10. ACM, New York, NY, USA, pp 393–402

Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58(7):1019–1031

Luhmann N (1993) Gesellschaftsstruktur und Semantik: Studien zur Wissenssoziologie der modernen Gesellschaft, vol 1. Suhrkamp Frankfurt/M

Markines B, Cattuto C, Menczer F, Benz D, Hotho A, Stumme G (2009) Evaluating similarity measures for emergent semantics of social tagging. In: Proceedings of 18th international world wide web conference (WWW'09), pp 641–650

Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296(5569):910

McGee J, Caverlee JA, Cheng Z (2011) A geographic study of tie strength in social media. In: Proceedings of 20th ACM international conference on information and knowledge management, CIKM '11, ACM, New York, NY, USA, pp 2333–2336

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27(1):415–444. doi:10.1146/annurev.soc.27.1.415

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27, pp 415–444 (2001). http://www.jstor.org/stable/2678628

Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, ACM, pp 29–42

Mitzlaff F, Atzmueller M, Benz D, Hotho A, Stumme G (2011) Community assessment using evidence networks. In: Atzmueller M, Hotho A, Chin A, Helic D (eds) Analysis of social media and ubiquitous data, LNAI, vol 6904. Springer, Heidelberg, Germany, pp 79–98

Mitzlaff F, Atzmueller M, Benz D, Hotho A, Stumme G (2013) User-relatedness and community structure in social interaction networks. CoRR/abs

Mitzlaff F, Benz D, Stumme G, Hotho A (2010) 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

Murata T, Moriyasu S (2007) Link prediction of social networks based on weighted proximity measures. In: Web Intelligence, IEEE/WIC/ACM international conference on, pp 85–88 IEEE

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E Stat Nonlinear Soft Matter Phys 69(2):1–15

Scellato S, Noulas A, Lambiotte R, Mascolo C (2011) Socio-spatial properties of online location-based social networks. In: Proceedings of the fifth international conference on weblogs and social media (ICWSM) vol 11, pp 329–336

Schifanella R, Barrat A, Cattuto C, Markines B, Menczer F (2010) Folks in folksonomies: social link prediction from shared metadata. In: Proceedings 3rd ACM international conference on web search and data mining, ACM, New York, NY, USA, pp 271–280

van de Rijt A, Kang SM, Restivo M, Patil A (2014) Field experiments of success-breeds-success dynamics. Proc Natl Acad Sci p 201316836

Yang J, Leskovec J (2011) Patterns of temporal variation in online media. In: Proceedings of the fourth ACM international conference on Web search and data mining, ACM, pp 177–186