The Social Distributional Hypothesis
A pragmatic proxy for homophily in online social networks

Folke Mitzlaff · Martin Atzmueller · Andreas Hotho · Gerd Stumme

Received: date / Accepted: date

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. These data collections comprise facts as well as subjective opinions about common concepts, but also new relations among concepts emerge. In linguistics, the Distributional Hypothesis [18] 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. This hypothesis has also been investigated in the context of social tagging systems [8, 31]. 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.

Keywords social networks · social interactions · social media · analysis · distributional semantics

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 relationships, e.g., by sending private messages and establishing friendship in Facebook. The hereby induced networks of user relatedness are a valuable source of information for different applications, considering, e.g., the task of recommending new acquaintances [9, 14] or finding groups of related users [1, 22, 41] for targeting a commercial campaign. But 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, by 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 [34], 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 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 [18]. For grounding this hypothesis, we consider a broad range of possible user interactions in Twit-
ter, Flickr and BibSonomy and analyze correlations between therefrom derived interaction characteristics and external metrics of user similarity. The remainder of the paper is structured as follows: Section 2 discusses related work. After that, Section 3 presents the necessary notions of the applied social network analysis methods. Furthermore, Section 4 formalizes the proposed social distributional hypothesis and presents the applied methodology for grounding the hypothesis. Next, Section 5 describes the applied datasets. The results of the experiments are described in Section 6. Finally, Section 7 concludes with a summary and interesting options for future research.

2 Related Work

Analyzing Web 2.0 data by applying complex network theory goes back to the analysis of (samples from) the web graph [7]. Mislove et al. [35] applied methods from social network analysis as well as complex network theory and analyzed large scale crawls from prominent social networking sites. Certain properties which are common to all considered social networks, are worked out and contrasted to properties of the web graph. Newman analyzed many real life networks, summing up characteristics of social networks [40]. 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 [25]. In [38], social interaction networks which are contracted using aggregations of log files within the social tagging system BibSonomy are introduced and analyzed. An analysis of a location-based social network with respect to user attributes is investigated in [27]. Interdependencies of social links and geospatial proximity are investigated in [21,33,45], especially concerning the correlation of the probability of friendship links and the geographic distance of the corresponding users. Silva et al. [48] mine structural correlation patterns in network partitions, i.e., correlations between vertex attributes and dense components in undirected graphs. While their approach results in individual patterns, our analysis captures both patterns and the networks/graphs as a whole and provides comprehensive analysis on their combined structure. Schifanella et al. [46] investigated the relationship of topological closeness (in terms of the length of shortest paths) with respect to the semantic similarity between the users. Crandall et al. [12] 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. Leroy et al. [26] discusses a feature-based approach using implicit information for inferring interaction networks in the context of link prediction. Eagle et al. [15] describe an approach for reconstructing friendship relations from secondary (mobile phone) data. They show, that friendship links can be inferred with a high probability but do not present a comprehensive analysis of different evidence networks and their impact on the predictability. Barrat et al. [4] discuss the relation between online and offline networks. Similarly, Chin et al. [10] consider networks of encounters for inferring contact networks, however, no relations to other evidence networks are discussed. Another aspect of our work is the analysis of implicit link structures which can be obtained in a running Web 2.0 system and how they relate to other existing link structures. Baeza-Yates et al. [3] propose to present query-logs as an implicit folksonomy where queries can be seen as tags associated to documents clicked by people making those queries. Based on this representation, the authors extracted semantic relations between queries from a query-click bipartite graph where nodes are queries and an edge between nodes exists when at least one equal URL has been clicked after submitting the query. Krause et al. [24] analyzed term-co-occurrence-networks in the logfiles of internet search systems. They showed that the exposed structure is similar to a folksonomy. Mitzlaff et al. [36, 37] analyze structural interrelations between evidence networks and provide evidences for strong ties between different networks. In addition, the work presents a novel method for community assessment. Another related area concerns link analytics, especially link prediction utilizing relations between temporal partitionings of a network. Most of these works analyzed the predictability of new links in online social networks like co-authorship in DBLP or arXiv.org. Schozl et al. combination of offline and online networks for link prediction [11]. Backstrom et al. [2] predict new links utilizing random walks. [50] presented prediction techniques using location-based proximity as a proxy for face-to-face encounters and online social networks. In contrast, [47] conducted a first analysis concerning the predictability of new links in real face-to-face contact networks. In the offline context, Oloritun et al. [43] investigate the interrelations between sensed interactions, closeness and shared places. In contrast to previous work, this paper 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 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., [13, 16, 42].

3.1 Graph & Network Theory

For modeling (social) networks, we use concepts and notions from the study of graphs, i.e., graph theory. In the following, we briefly introduce basic notions used in this work and refer to standard literature, e.g., [13], for a detailed introduction and discussion of graph theory.

Basic Concepts and Notation An undirected graph G = (V, E) consists of a finite set V of vertices or nodes, and a set E of edges, which are two element subsets of V. In a directed graph, E denotes a subset of V × V. For simplicity, we write (u,v) ∈ E in both cases for an edge belonging to E and freely use the term network as a synonym for a graph. The density of G is the fraction of possible edges that are actually present. In a weighted graph, each edge l ∈ E is given an edge weight w(l) by some weighting function w : E → ℝ. The degree of a node in a network is the number of connections it has to other nodes. The adjacency matrix of a set of nodes S with n = |S| contained in a (weighted) graph G = (V, E) is a matrix A ∈ ℝ^{n×n} with 

A_{ij} = \begin{cases} w(i,j) & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}

for any nodes i, j in S (assuming some bijective mapping from 1, ..., n to S). We identify a graph with its adjacency matrix and freely use the term adjacency matrix.

For an undirected graph G = (V, E) of minimum length and the diameter of G is the shortest path of maximal length. The transitive closure of a graph G = (V, E) is given by G* = (V, E*) with (u,v) ∈ E* iff there exists a path u →_G v. A strongly connected component (scc) of G is a subset U ⊆ V, such that u →_G v exists for every u, v ∈ U. A (weakly) connected component (wcc) is defined accordingly, ignoring the direction of edges (u,v) ∈ E. A binary relation on a set V is a relation R as a subset R ⊆ V × 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) ∈ 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) ∈ 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. Implicit relations can be derived or constructed by analyzing secondary data. Many observations of network properties can be explained just by the network’s degree distribution [23]. 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 [32].

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 [28]. In the following, we present all considered similarity functions, following the presentation given in [44] which builds on the extensions of standard similarity functions for weighted networks from [39]. The Jaccard coefficient measures the fraction of common neighbors:

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

The Jaccard coefficient is broadly applicable and commonly used for various data mining tasks. For weighted networks the Jaccard coefficient becomes

\[ \overline{JAC}(x,y) := \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(y,z)}{\sqrt{w(x,z)} \sqrt{w(y,z)}} \]

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

\[ COS(x,y) := \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| \cdot |\Gamma(y)|} \]

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

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

Preferential PageRank (PPR) The preferential (or preference) PageRank similarity is based on the well known PageRank [6] algorithm. For an \( m \times m \) column stochastic adjacency matrix A and damping factor α, and uniform preference vector \( \mathbf{p} := (1/m, \ldots, 1/m) \), the global PageRank
vector $w =: \text{PR}$ is given as the fixpoint of the following equation:

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

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 $\text{PPR}(\mathcal{I})$ 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 call therefrom aggregated networks of user relatedness “evidence networks”, in reference to [19], 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 [34], 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 Figure 1b shows an evidence network derived from Twitter’s Follower graph, Figure 1c shows a different evidence network for the same set of users, which is derived from Twitter’s ReTweet graph (cf. Sec. 5.1).

4 The Social Distributional Hypothesis

This work aims at grounding the social distributional hypothesis which states that users with similar interaction characteristics in applications of the social web tend to be related. In contrast to linguistic entities, such as words [18] or tags [8, 31], which are associated with certain semantics (e.g., the thing denoted by a word), users lack of 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 [30]). 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 and analyzing their interdependence with interaction characteristics of users in social web applications. In particular, we consider geographic proximity and similarity of the applied tag vocabulary of different users. More detailed, we consider the following three research questions:

1. Are people who interact more frequently tend to be more similar?
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?

We conduct a series of according experiments on social interaction networks (cf. Sec. 3.3), derived from different typical social web applications. Corresponding to the first research question, we consider the average pairwise covariate similarity of users (e.g., the 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 $\text{COS}(u, v)$, $\text{JAC}(u, v)$, $\text{PPR}(u, v)$ with variants (cf. Sec. 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 $\hat{G}$ (cf. Sec. 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 [46]. As for the second research question, 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. In order 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). Concerning 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 [28] and distributional semantics [20, 31]. 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. Sec. 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 hashtag 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.

5 Data

For conducting our experiments, we aggregated various explicit and implicit evidence networks of user relatedness 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 firstly describe the considered evidence networks and summarize corresponding general high level statistics (these networks are thoroughly analyzed in a comparison [37]). Then, we present the collected data sets of external user properties.

5.1 Network Data

We consider networks of users, where edges between users are used to model relations. Depending on the nature of the underlying interactions, edges are accordingly directed or undirected and modeled with or without edge weights.

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 [49], 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 [32] 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.

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 es-
tablishing 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 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 towards high degree nodes \(^1\) and likely underestimate link symmetry \(^5\). 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 crawled in parallel. Thus, induced biases have a comparable impact on all considered networks.

**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, less than 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.

**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

---

\(^1\) http://data.gov.au/1277
\(^2\) http://www.flickr.com/services/api/
\(^3\) http://delicious.com/network/<username>
\(^4\) http://www.bibsonomy.org/friends
\(^5\) Note: For privacy reasons a user may deactivate this feature.
In the context of social tagging systems like BibSonomy, the cosine similarity is often used for measuring semantic relatedness [8, 31]. We compute the cosine similarity in the vector space $\mathbb{R}^T$, where for user $u$, the entries of the vector $u := (u_1, \ldots, u_T) \in \mathbb{R}^T$ 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 $u$ and $v$ (cf. Sec. 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 [31].

### 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 ten-

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.

#### 5.2 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. Sec. 5.1). We also consider geographical distance of users in Twitter and Flickr.

**Tag Based Similarity** In the context of social tagging systems like BibSonomy, the cosine similarity is often used for measuring semantic relatedness [8, 31]. We compute the cosine similarity in the vector space $\mathbb{R}^T$, where for user $u$, the

| network | $|V|$ | $|E|$ | density | #SCC | ISCC | WCC |
|---------|------|------|---------|------|------|------|
| Copy    | 1,427| 4,114| $2 \cdot 10^{-3}$ | 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| $2 \cdot 10^{-3}$| $-$  | $-$  | 228  |
| Friend  | 700  | 1,012| $2 \cdot 10^{-3}$| 515  | 17   | 238  |
| ReTweet | 826,104 | 2,286,416 | $3.4 \cdot 10^{-6}$ | 699,967 | 123,955 | 702,809 |
| Follower| 1,486,403 | 72,590,619 | $3.3 \cdot 10^{-5}$ | 198,883 | 1,284,201 | 1,485,356 |
| Comment | 525,902 | 3,817,626 | $1.4 \cdot 10^{-5}$ | 472,232 | 53,359 | 522,212 |
| Favorite| 1,381,812 | 20,206,779 | $1.1 \cdot 10^{-5}$ | 1,305,350 | 76,423 | 1,380,906 |
| Contact | 5,542,705 | 119,061,843 | $3.9 \cdot 10^{-6}$ | 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.
dencies can be observed by averaging over many observations. We used Yahoo!'s Placemaker™ API⁶ 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. The geographical distance of users is then simply given by the distance of the centroids for the correspondingly matched places. Please note that geographical distance correlates with many secondary notions of relatedness between users, such as, e.g., language, cultural background and habits.

6 Experiments

Within this section, we present the results of the experiments from Sec. 4 as applied on the different networks and semantically grounded user similarity metrics described in Sec. 5. 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.

6.1 Grounding of Interaction Proximity

Considering our first research question, whether interacting users tend to be more similar then those who do not interact directly, we investigate, more generally, whether a negative correlation between the average pairwise semantic similarity and the corresponding shortest path distance of users in evidence networks can be observed (cf. Sec. 4).

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.

Geographic Distance For average geographic distances of users in Flickr and Twitter, we repeated the same calculations, and show the obtained results in Figure 4. Firstly, 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 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 be-

---

⁶ http://developer.yahoo.com/geo/placemaker/ (November 2011)
between distance in the interaction network and geographical distance.

Discussion It is worth emphasizing that in all considered evidence networks, the relative position of users already 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 at 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.

6.2 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 (cf. Sec. 4). In order 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 \( c := 10^{-8} \), resulting in \( b := c^{-1} \). 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.

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 Twit-
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.

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 distance drops below 2,000 kilometers, in contrast to the global average pairwise distance of 7,400 kilometers. 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.

Discussion 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. But also the very implicit interaction of visiting a user’s profile page in BibSonomy already gives rise to tendencies of user interrelationship. Altogether, the observed tendencies of higher similarity scores for increasing interaction frequencies give evidence for the postulated social distributional hypothesis, that interacting users tend to be semantically related.

6.3 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 towards 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 (cf. Sec. 4). We firstly consider tag based similarity of users in BibSonomy, Flickr and Twitter and then geographical 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 towards lower similarity scores (most pairs of nodes have very low similarity scores), we applied logarithmic binning (cf. Sec. 6.2).

**Semantic Similarity** In Figure 7, the obtained results are shown 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 \( > 10^{-4} \).

**Geographic Distance** As for geographic distances, Figure 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 \( \text{COS} \) and \( \text{JAC} \), 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 \( \text{COS} \) and \( \text{JAC} \) 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 \( [0, 10^{-4}] \)), but then increasing distances for higher similarity scores.

**Discussion** The obtained results point at tendencies of the considered similarity metrics in capturing tag based semantics similarity and geographic proximity of users by means of structural similarity. In most cases, the preferential PageRank similarity best captured geographic proximity of users. 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 thereby seems to encompass the strongest geographic binding for direct neighbors, as indicated in the relative low average distance for direct neighbors (cf. Fig. 4).

7 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.

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

References