Online Social Networks and Media 2 (2017) 1–18

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Semantic homophily in online communication: Evidence from Twitter

Sanja Šćepanović[a,*], Igor Mishkovski[b], Bruno Gonçalves[c], Trung Hieu Nguyen[d], Pan Hui[e,f]

[a] Aalto University, Department of Computer Science, Espoo 02150, Finland
[b] University Ss. Cyril and Methodius, Faculty of Computer Science, Skopje 1000, Macedonia
[c] New York University, Center for Data Science, New York, NY 10011, USA
[d] University of Tampere, Faculty of Natural Sciences, Tampere 33100, Finland
[e] Hong Kong University of Science and Technology, Department of Computer Science and Engineering, Clear Water Bay, Kowloon, Hong Kong
[f] University of Helsinki, Department of Computer Science, Helsinki 00560, Finland

ARTICLE INFO

Article history:
Received 2 March 2017
Revised 12 June 2017
Accepted 26 June 2017

Keywords:
Homophily
Semantics
Influence
Semantic relatedness
Twitter
Wikipedia
Social network analysis
Computational social science

ABSTRACT

People are observed to assortatively connect on a set of traits. This phenomenon, termed assortative mixing or sometimes homophily, can be quantified through assortativity coefficient in social networks. Uncovering the exact causes of strong assortative mixing found in social networks has been a research challenge. Among the main suggested causes from sociology are the tendency of similar individuals to connect (often itself referred as homophily) and the social influence among already connected individuals. Distinguishing between these tendencies and other plausible causes and quantifying their contribution to the amount of assortative mixing has been a difficult task, and proven not even possible from observational data. However, another task of similar importance to researchers and in practice can be tackled, as we present here: understanding the exact mechanisms of interplay between these tendencies and the underlying social network structure. Namely, in addition to the mentioned assortativity coefficient, there are several other static and temporal network properties and substructures that can be linked to the tendencies of homophily and social influence in the social network and we herein investigate those.

Concretely, we tackle a computer-mediated *communication network* (based on Twitter mentions) and a particular type of assortative mixing that can be inferred from the semantic features of communication content that we term *semantic homophily*. Our work, to the best of our knowledge, is the first to offer an in-depth analysis on semantic homophily in a communication network and the interplay between them. We quantify diverse levels of semantic homophily, identify the semantic aspects that are the drivers of observed homophily and show insights in its temporal evolution. By analyzing these mechanisms we increase understanding on what are the semantic aspects that shape and how they shape the human computer-mediated communication. In addition, our analysis framework presented on this concrete case can be easily adapted, extended and applied on other type of social networks and for different types of homophily.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Homophily [17,35,37,44] represents the tendency of individuals who are similar on some traits to connect (become friends, follow each other, communicate etc.) in a social network. **Social influence** (peer pressure) is the inverse tendency of becoming similar on some traits or adopting certain behavior from your social contacts. Both, homophily, and social influence result in assortative mixing in the network. *Assortative mixing* or social correlation means that neighbors exhibit a higher similarity compared to random users in the network. Assortative mixing is repeatedly confirmed in social network analysis literature [3,5,11,20,70]. A question remains, to what extent is the observed assortative mixing a result of an underlying homophily that shapes the formation of the network or of the social influence taking place in an already formed network [37]. In addition to individuals connecting to similar individuals, **link dissolution** over time among non-similar individuals is another suggested mechanism relating to homophily. Yet another factor that could cause assortative mixing is a common **external influence**. Moreover, a combination of these factors is often at play. For instance, an external factor might have a non-homogeneous adoption in the network because friends have a higher common latent propensity for it and adopt it to a larger extent than non-friends. Distinguishing between these factors as the causes of assortative mixing has been a challenge, and proven not even possible from observational data [68].

* Corresponding author.
E-mail address: sanja.scepanovic@aalto.fi (S. Šćepanović).

http://dx.doi.org/10.1016/j.osnem.2017.06.001
2468-6964/© 2017 Elsevier B.V. All rights reserved.

Extensive research is conducted in sociology on homophily in social networks (see the seminal review by McPherson et al. [44]). The classical paper [35] from the middle of the 20th century introduced two basic levels of homophily: status and value homophily. **Status homophily** relates to any formal or perceived status among individuals. It includes some of the most important social dimensions, such as *race, ethnicity, sex, age, education, occupation* and *religion*. **Value homophily** relates to our internal states that might shape the future behavior; for example: *abilities (intelligence), aspirations*, and *attitudes (political orientation)*, regardless of the differences in status.

The picture on homophily is additionally complicated because not all social links are equal. In his famous Strength of Weak Ties (SWT) theory, Granovetter [28] discussed the differences between **strong links** and **weak links**. He defined strong links through a combination of time, intimacy, emotional intensity and reciprocity. Granovetter also argued that users connected with a strong link are more likely to be similar. Weak links correspond to acquaintances with whom there is usually less interaction. However, those are the bridging links that introduce the node to novel and potentially useful information. Another similar theory introduces the advantageous network position of bridging nodes using the concept of **structural holes** [13]. Burt [13] argues that the nodes positioned between different network groups are more familiar with alternative ideas and behaviors compared to the nodes embedded in dense clusters. He developed a measure of structural holes corresponding to the advantageous social position that is named *Burt's index* after him. Described *network position* can be a more implicit cause of homophily. Research finding exist that holding a same position inside an organization will induce larger homophily between individuals than it would be the case if the links were random [42].

1.1. Background and motivation

With online social networks (OSNs) becoming widely used, researchers are nowadays gaining access to data that were previously unimaginable in sociology. High numbers of OSN users and their active engagement create large-scale datasets of social interactions. This in turn has lead to a wave of new research tackling introduced and many other sociological phenomena within OSNs [22,80]. Such research lies in between sociology, on one side, and computer science required for processing and analyzing large-scale datasets, on another. Hence a new research field termed *computational social science* emerged [36]. Our study is situated within this field. We study introduced tendencies of status and value homophily, their interplay with links of different strength and through time in the context of a particular OSN, Twitter.

Depending on the purpose of OSN, connections among its users describe a different relationship. In most of the studies on Facebook, Twitter, Instagram or Flickr, for instance, connections are taken to represent friendship, acquaintanceship or followership. Such links are *binary* (connected or not) and *one time* (once formed and persist). Some other binary and one time relationships are citations (bloggers) or professional cooperation (LinkedIn). However, almost all OSNs provide additional layers of *interactions* that enable to assign other types of links to the users. For instance, users can send messages, write wall posts, they can mention each other or in some networks endorse or give rating. Such interactions enable to assign **strength** and **temporal dimension** to the links. In our study on Twitter data we analyze underlying communication via direct messages as the basis for forming links. This makes our study different from a number of previous studies in which assigned links are binary and one time. Xiang et al. [76] developed a connection strength indicator based on interaction activity and argued how it offers more information compared to binary links. While not developing a sophisticated model as Xiang et al. [76], we also consider the links between users based on their interaction frequency (number of mentions) instead of binary links. The frequency and times of interaction add a temporal dimension: links in our network can be persisting, or they can form, decommission and form again during the dataset time-frame.

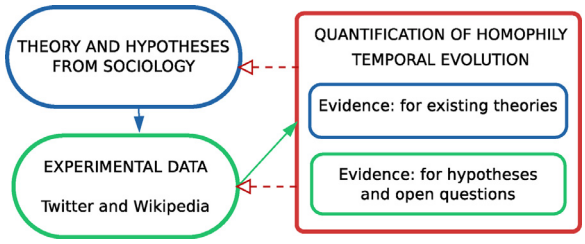

Fig. 1. General framework of our study. In addition to the evidence found for some existing sociological theories in the empirical Twitter data, we also find evidence pointing to some hypotheses. However, such hypotheses remain open questions as they should be evaluated and confirmed in several other datasets before any general conclusions about semantic homophily in communication are reached.

Fig. 1 presents the general framework of our study. As depicted, we operate on empirical datasets (from Twitter and Wikipedia), while basing our research questions on existing sociological theories. **Communication network** that we analyze is built from Twitter mentions. A natural method to asses homophily in communication is through semantic aspects of it, hence we talk about semantic homophily. We define **semantic homophily** as the degree of assortative mixing on semantic aspects in communication. We build a semantic database from a Wikipedia dataset according to the Explicit Semantic Analysis (ESA) [25] algorithm. This semantic database and the IBM Alchemy API [2] enable to assess semantic aspects of tweets. Through these semantic aspects, in turn, we measure semantic homophily.

In the following section, we summarize the terminology used in this study.

1.2. Terminology

Communication network is based on Twitter mentions, representing computer-mediated online communication [73]. The network nodes correspond to Twitter users, and the directed links to the tweets in which they mention each other. The tweet content is also included. Our network is a subtype of previously introduced *interaction networks* [74]. The links in our network are not only formed once (one time), but they require an active engagement over time. They are also not binary, but have strength based on the number of mentions. When a user A follows a user B it simply states some type of potential interest in what B has to say. Depending on their time zones and on the number of other users that A is following, she might not even get to see any of B's tweets. In the case of our communication network we can clearly point to an interaction and information diffusion (when the user A mentions the user B), instead of simply speculating about it when using the friendship/follower network. Communication generally refers to exchanging of information. We recognize the potential of Twitter mentions to carry two different forms of information exchange. In the first form, the source is directly addressing the receiver. In the second form, there is a sort of authority attribution where the source comments to the rest of the Twitter users about the receiver (this could be a critique as well).

Communication intensity (CI) denotes the weight on the links i.e., the number of mentions between a pair of users. CI is also taken as an indicator of **link strength** after SWT theory [28].

Assortative mixing or **social correlation** represents higher similarity on certain traits between neighbors compared to random users in a network. In other words, it is preferential attachment

of individuals who are similar on some traits. Given a network, assortative mixing can be measured through several assortativity coefficients (see Methods 7.1).

Semantic homophily is assortative mixing on semantic aspects of communication. We point out that in many related studies, the term *homophily* is used with the meaning of *assortative mixing* as we introduced it here. A possible reason is the described indistinguishability of homophily with social influence. We also use the term semantic homophily, while a more precise term would be *semantic assortative mixing*. However, we select to talk about semantic homophily in order to be consistent with the related studies and also since we do not focus on distinguishing between homophily and social influence. In order to measure semantic homophily, we tackle several semantic aspects of communication discussed in the following.

Semantic relatedness (SR) represents similarity and relatedness of topics in two texts. SR is a more general metric compared to semantic similarity [30] since in addition to similarity, it includes also any other relation between the terms, such as antonymes (opposite terms) [38] and meronymes (a term is a part of or member of the other) [48]. For instance, the term *airplane* is similar to the term *spacecraft*. The same term is related to *car, train* or *wing*, but not similar to them. SR between texts is quantified by a value ranging from 0 (not related at all) to 1 (maximally related).

Sentiment of a text receives a value ranging from −1 (negative) to 1 (positive).

Entities represent concrete topics discussed in a text, such as: people, companies, organizations, cities and geographic features.

Concepts represent abstract ideas in a text. *For example, if an article mentions CERN and the Higgs boson, it will have Large Hadron Collider as a concept even if the term is not mentioned explicitly in the page* [2].

Communication propensity ($\hat{cp}$) is a function of some network property measuring the extent to which the observed communication and its intensity diverge from what would be expected in a uniformly random setting with respect to that property. We investigate communication propensity with respect to SR threshold (formula in Section 4.1).

Social capital [12,59] represents actual and potential resources that are linked to the ego's social network and relationships. Hence, in similarity to the socio-semantic networks [65], we define social capital in our communication network as the total *number of contacts* (degree in the unweighted network) or the total *strength of contacts* (degree in the weighted network). Moreover, in both cases we can separate *popularity* (indegree) and *activity* (outdegree). Another way we use to access social capital is in terms of network position measured with *Burt's index* [13] (see Methods 7.2).

Semantic capital denotes the amount of diversity in a text with respect to the above introduced semantic attributes, similarly as in [65].

Relative status of two users represents the difference of their respective status ranks. It can be defined for both social and semantic capital. Relative status is termed achieved status in some studies [69].

User status inconsistency [39,62] is a relative difference between her ranks among all users on social and semantic capital. Status inconsistent individuals tend to be highly ranked on some aspects and lowly ranked on others.

Link status inconsistency measures total user status inconsistency of the communicating users (we give a formal definition in Section 4.4).

1.3. Contributions

The focus of our work is on **semantic homophily**. While several other studies have tackled some aspects of semantic homophily, as we discuss in the related work, to the best of our knowledge this is the first study aiming towards a comprehensive picture on the role of semantic homophily in communication. We offer an in-depth and detailed investigation of semantic homophily: from quantification and qualitative assessment to its temporal evolution. Below we summarize our findings in the context of the framework in Fig. 1:

- **Quantification of homophily**
  - We confirm that semantic relatedness of tweet contents is a homophilous trait for communication. Using communication propensity, we measure the extent of the homophily. Additionally, there are differences between strong and weak links. While similarity increases the probability of forming strong links, the weak links are almost equally likely across user pairs with different similarities. *Hence we confirm a specific aspect of the SWT [28] in communication for semantic homophily.*
  - Granovetter's theory also suggests that, with similarity, the link strength between already connected users is higher [28]. *We confirm that the intensity of communication as a proxy for the link strength correlates with user pair similarity.* However, we do not establish any causal relationship. Moreover, the correlation is rather moderate leaving space for many other factors that affect the strength of the links in addition to semantic similarity that we measured.
  - We formulate several forms of social and semantic capital [12] and measure respective status and value homophily levels [35] in communication. Forms of social capital are operationalized through network position based on degree, strength of connections and Burt's index [13]. Forms of semantic capital are defined through sentiment and the diversity of the user tweets with respect to entities and concepts.
  - Social status homophily is manifested in assortative degree mixing in the network (degree assortativity) [51]. We calculate different degree assortativity coefficients (undirected and directed, unweighted and weighted) and show that Twitter mention network is not assortative for some of them when we consider also the weak links. However, as we shift the focus towards strong links, the network becomes strongly assortative on all the degree assortativity coefficients. These results exhibit high social status homophily in communication, particularly for strong contacts. *In this way we also reveal important differences between communication network captured by mentions on Twitter and previously analyzed follower type of network.*
  - The interplay between social and semantic capitals reveals that popularity and activity correlate positively with semantic diversity, and negatively with sentiment. On the other hand, social capital measured through Burt's index shows no correlation with either of the forms of semantic capital. *In other words, the advantageous position in the communication network is not due to any semantic superiority of the users.*
  - Finally, we operationalize sociological concept of status inconsistency [62] in the Twitter network and show that link inconsistency correlates with its strength, i.e., communication intensity.
- **Temporal evolution**
  - Building on several previous studies [17,78] that investigated homophily during link formation, to the best of our knowledge, we are the first to *investigate its change also during link decommission*. Our results reveal the role of status and value homophily during link formation and decommission.
  - Firstly, as reported in the previous studies, we also detect an increase in user semantic relatedness (value level homophily) prior to link formation. Then we also observe

a drop in similarity between users once they become connected.
- We go a step further, by separating persisting from the less stable links, to exhibit that it is due to the later that the average similarity drops. However, for the pairs who actively persist communication, the similarity stays high also after their link formation.
- In relation to status level homophily, we find that pairs who start communicating and have similar social status are more likely to persist compared to pairs who start with a larger difference in status.
- Respectively, *the pairs who eventually cease communication are not characterized by semantic differences (value heterophily), but by status differences (status heterophily).* This is evidence in support of sociological theories that links with dissimilar individuals are more expensive and difficult to maintain [10,24]. However, we find this evidence mainly for the status level similarity.

The rest of the paper is organized as follows. Section 2 presents related research literature. In Section 3 we describe the two datasets (from Twitter and Wikipedia), as well as the framework for analysis consisting of a communication (Section 3.1) and semantic (Section 3.2) layer. Quantification of different forms of homophily is presented in Section 4: the relationship between semantic relatedness and communication is reported in Section 4.1, social status homophily in Section 4.2 and semantic status and value homophily in Section 4.3. Insights on the relationship between different forms of capital are given in Section 4.4, insights on relative status in Section 4.5 and on status inconsistency in Section 4.6. Temporal aspects of semantic homophily, including link formation, dissolution and persisting interactions, are discussed in Section 5. The article concludes with a discussion and remarks on future research in Section 6.

2. Related work

While homophily has been identified in a diverse set of social networks, most of the studies investigated friendship, followers or citation type of links. *Interaction links* are more suitable for inferring meaningful social relationships [74]. An interesting setting is presented in a study on Flickr data [27], where the followers layer and the interaction layer (photo favorites) are brought together in a multiplex network. The nature of the Twitter mention network (that is our focus) is fundamentally different from follower/friendship network in Twitter [9]. For instance, the reciprocity of the followers network is found to be around 22% [33] which is lower compared to the other social networks. The reciprocity of our mention network is 64%. While retweet links allow for a similar information diffusion analysis, their nature is also shown to be importantly different from mention links [15].

A similar study to ours [65] investigated knowledge networks representing scientific collaborations and blogger citations. The focus on the joint dynamics and co-evolution of the *socio-semantic structures* is similar to ours. However, we focus on another type of a network (communication). Hence, we respond in part to the call by Roth and Cointet [65] to analyze some of the epistemic patterns, which they found in the scientist and blogger communities, in other type of communities. Moreover, while they only investigate social link formation, we also investigate *link decommission*. Therefore, our work offers an additional understanding on the temporal interplay between semantic and social structures. Another important difference is that we offer a considerably deeper analysis of semantic aspects. Compared to a hand-picked set of categories used in [65], our Wikipedia-based database in combination with Alchemy API provide us with richer insights on entities, categories, taxonomy and also sentiment of communication. A recent study on Twitter analyzed homophily on the status level (defined as the difference in the follower counts) and the value level (tweet contents, common followees, location, age etc.) [69]. The focus on reciprocal followers network (instead of mention network in our case) and the analysis of homophily on the organization-individual relationship (whereas we focus on individual-individual relationship) are the main differences to our study. Another study on Twitter quantified purity homophily (sharing moral values) [21] using semantic analysis. Again, the main difference is the focus on follower network. Also, while they look at the particular type of value homophily (purity) through semantic aspects of tweets, we investigate the semantic homophily in general.

Several previous studies investigated *temporal interplay between homophily and social links* in online settings. Crandal et al. [17] found that the similarity between two Wikipedia admin users sharply rises some time before the link formation and after that continues to slowly grow. This is interpreted so that, at first, homophily plays a role in the link formation, but after that, the link plays a role in the continuous increase of homophily. Another similar study on Flickr [78], finds more subtle insights: the users who have similar popularity (defined as the average number of favorites for their photos) are more likely to diverge in similarity after the link formation; while the similarity continues to grow for the users who have a larger popularity difference. This is explained by the tendency of users to stay unique and diverse in their uploaded content from equally popular users. Through focus on a different type of social links (communication), and on a different form of homophily (semantic), our work extends these previous studies. Importantly, we add the insights about homophily during *link decommission* that they have not investigated. Moreover, we also uncover the relationship between introduced social and semantic forms of capital and homophily around the time of link formation and decommission. Link formation, only in relation to the network position and not to homophily on value level, is analyzed in [1]. A theoretical network model which includes homophily, link formation and decommission is presented in [79].

Halberstam et al. [29] analyze communication on Twitter (comprising both retweets and mentions of political candidates) in similarity to us, however, with a different aim – to understand information diffusion. They find a greater degree of homophily exhibited and also more connections per node in larger communities. A number of studies are conducted toward *distinguishing between influence and homophily* [3,4,34] reporting different levels and proportions of the two traits in online social networks. For example, De Choudhury et al. [20] quantified the impact of various types of homophily on influence on Twitter. Users were given homophilous traits based on attributes such as: location, information roles they take (generators, mediators and receptors), content creation (meformer, informer) and activity behavior (number of tweets per period of time). However, current methods for identifying causal effects are limited [14] and homophily and influence are indistinguishable in empirical settings due to confounding effects [68]. A couple of more recent papers tackled this research challenge in controlled experiments. The experiment on Facebook found that the probability for a user to share a link increases with the number of friends who shared the same link even without the user being exposed to their link shares [7]. Hence this controlled experiment confirmed homophily or some unobserved common external influence taking place in the network. Generative models to explain homophily and segregation patterns are explored in [18,19] and the evolution of OSN structures with regard to users' historical preferences in [75].

Table 1
Twitter dataset filtering steps statistics.

Dataset	Mentions	Users
Original download	12,441,636	547,368
English language	2,527,990	284,100
Users > 20 tweets	1344,692	29,616
Internal replies	**744,821**	**26,717**

Table 2
`Communication network` statistics.

Network parameter	Value	Network parameter	Value
Nodes	26,717	Max out-degree	1358
Edges	99,910	Max in-degree	3228
Avg weighted deg.	55.75	Diameter	29
Avg clustering coeff.	0.051	Density	0.00014

3. Datasets and analysis framework

3.1. Communication layer: Twitter mention network

Our initial dataset contains $12,441,636$ mentions (tweets including @username) among $547,368$ users over the course of 6 months (May–Nov 2011). *All internal mentions* are included, meaning, each time when a user from our dataset mentions a user from outside, we did not keep such tweets, but all the mentions among the users in the dataset are present.

In order to have a well suited dataset for the intended analysis, we perform several cleaning and filtering steps described below. The initial dataset includes tweets in several languages, so we filter it to select only English tweets and from the users who mostly tweet in English. We use NLTK Python library [8] in this step. After the language filtering, the dataset is reduced to 20% of its original size in terms of tweets, while the number of users halved. For semantic analysis, individual tweets are often too small and noisy, so the next step involves filtering the remaining users based on their total number of tweets. Upon research and pre-test with the semantic database that we built (described in the following section), a threshold of minimum 20 tweets is selected. After this step, the dataset contains $29,616$ users. Finally, again keeping only the internal replies withing this group of users, we end up with $26,717$ users in the dataset used to build the communication network (see Table 1).

In communication network, $G = (V, E, W)$, the nodes u_i, $u_j \in V$ represent Twitter users; they are connected with a directed edge $e_{ij} = (u_i, u_j) \in E$ if a user u_i mentions u_j, and the edge is assigned the weight $w_{ij} = (u_i, u_j) \in W$ equal to the communication intensity (total number of such mentions). Properties of the communication network are given in Table 2.

3.2. Semantic layers: semantic enrichment of communication network

On top of the communication layer, we extract another, *semantic layer* from the Twitter data. We apply two semantic analysis procedures that enrich our communication network in terms of *node* and *edge attributes*. The first procedure is based on **Wikipedia semantic database** that is built by ourselves (see following Section 3.2.1 for an overview and Methods 7.3 for details). The second procedure employs an existing, natural language processing API, **AlchemyAPI** [2] from IBM (explained in Section 3.2.2). The semantic database provides enrichment for both, links (SR between tweets of two users) and nodes (extracted Wikipedia concepts relevant to the user tweets – see following paragraph for details). AlchemyAPI provides an additional set of node attributes: concepts, entities, taxonomy and sentiment of the user tweets.

3.2.1. Wikipedia semantic database (details in Methods 7.3)

The semantic database can be queried for SR between two tweet collections, i.e., between two users. This gives us *link attributes* in the semantic layer. In a somewhat computationally demanding task, we calculate the SR scores between *all the user pairs* (not just those who communicate and are connected in the communication network), resulting in the full **SR network**. During the analysis, we also apply different thresholds (SR_{th}) on the edge weights and obtain several SR sub-networks, which we denote **SR_th networks**.

In addition to SR scores, for each user we can also obtain their corresponding Wikipedia **concept vectors** *CV*s. *CV*s are formed of relevant Wikipedia concepts (articles) semantically describing the user tweet contents. This gives us *node attributes* in the semantic layer.

3.2.2. AlchemyAPI

AlchemyAPI (later merged with of IBM Watson Developer Cloud[1]) [2] performs natural language processing (NLP) and machine learning (ML) analysis. We send individual user tweets collections for analysis and AlchemyAPI returns semantic meta-data from the content. Not all are relevant for our study but we utilize following: sentiment score, taxonomy, concepts, entities and keywords. Hence, based on the output, we assign a set of attributes to users: the overall sentiment of her tweets (a real number between -1 for fully negative and 1 for fully positive), the taxonomy hierarchy representing topics, concepts, entities and keywords found relevant in the tweets. For each of the elements in the output, AlchemyAPI also returns corresponding relevance score, that we utilize to filter for most relevant semantic attributes.

Based on the evaluations in the literature, we believe that AlchemyAPI is a suitable choice to support our work. In [45] it was shown that the sentiment analysis obtained from AlchemyAPI achieved accuracy of 86% on a corpus of 5370 tweets employed by an intelligent recommendation system for tourism. The AlchemyAPI's performance on a number of datasets and in different contexts was also shown in [61] and [66], where AlchemyAPI outperfomed Zemanta,[2] OpenCalais,[3] Extractiv[4] and DBpedia Spotlight[5] in extracting and categorizing named entities. However, besides the evaluations stated above, and the benchmark analysis done in [60], we might consider using sentence-level methods, as VADER [32], SentiStrength [72] or Umigon [41] on our Twitter dataset as our future work.

4. Quantifying semantic homophily

4.1. Semantic relatedness as a homophilous trait for communication

To what extent are semantically related users more likely to communicate? In other words, how strong is the homophily in terms of semantic relatedness in communication? This question can be answered using communication propensity. **Communication propensity** with respect to SR ($\hat{cp}(SR_{th})$) is the extent to which observed communication for user pairs with relatedness higher than some threshold SR_{th} diverges from communication expected of random user pairs. To illustrate, SR_0.2 network has $\sim 40M$ links, or $\sim 9\%$ of all the possible $\sim 438M$ links in full SR network. Hence, in a uniformly random setting, we would expect a similar percent of communication links in SR_0.2 network. However, we find this percent to be 3 times higher. Precisely, the dyadic

[1] https://www.ibm.com/watson/developercloud/doc/index.html
[2] http://blog.zemanta.com/
[3] http://www.opencalais.com/
[4] http://extractiv.com/
[5] https://github.com/dbpedia-spotlight/dbpedia-spotlight

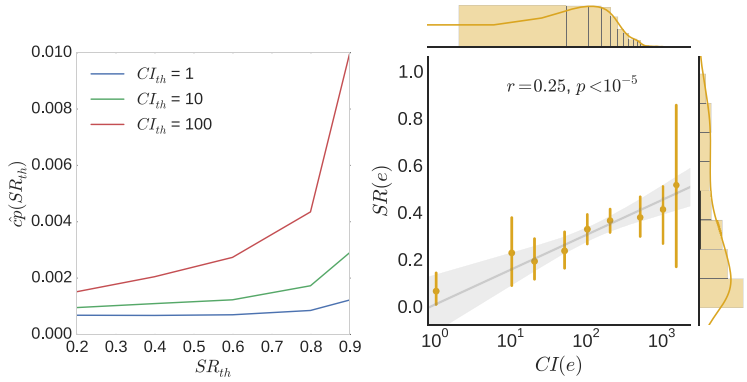

Fig. 2. SR as homophilous trait in communication: (left) $\hat{cp}(SR_{th})$ communication propensity with respect to SR for links with different minimum strengths based on communication intensity CI_{th}; (right) correlation of link's e SR and its strength based on CI. To account for long-tailed distribution of $CI(e)$ the we apply logarithmic binning [46] in visualization: average value and standard deviation are shown for each bin.

propensity formula defined in [64] can be used to calculate $\hat{cp}$:

$$\hat{cp}(SR_{th}) = L_{comm}(SR_{th})/L_{tot}(SR_{th}),$$

where $L_{comm}(SR_{th})$ is the number of links in communication network with SR value higher than the threshold and $L_{tot}(SR_{th})$ is the number of total possible such links.

With this approach we can also calculate propensity for links of a minimum given strength based on CI_{th}. Fig. 2 (left) presents results for weak links $CI_{th} = 1$ and for two levels of stronger links. Communication propensity for weak links increases only slightly for higher SR values, while for stronger links it starts to grow faster and reaches an order of magnitude higher values for $SR = 0.9$ and $CI_{th} = 100$. The results confirm that semantic similarity is a homophilous trait in communication, however much more so in the case of strong than weak links. *This result is in agreement with Granovetter's theory* [28].

A similar, but subtly different question can be asked for already established links. Are stronger links characterized by higher semantic relatedness of the communicating users? Fig. 2 (right) visualizes the relationship between communication intensity CI, as a proxy for link strength, and SR. Moderate statistically significant correlation is present. However, we find that user pairs exist who communicate quite intensively but have low relatedness of their tweet contents and also on the opposite – some users with relatedness close to 1 seldom communicate. Similar results are found in a study that evaluated this type of relationship in retweet and follower Twitter graphs [47].

4.2. Social status homophily

As we introduced earlier, a basic measure of assortative mixing in a network is the assortativity coefficient [50] or simply assortativity. This coefficient is a Pearson correlation between the properties of linked nodes. For the formulas used in our calculations, see Methods 7.1.

4.2.1. Undirected network variants

Positive degree assortativity [52] is suggested to be fundamental to social networks and to distinguish them from other types of networks [51]. We start by looking at the undirected variant of the communication network. Such an abstraction provides us with social capital in terms of number (unweighted) and strength (weighted) of contacts.

The coefficients r calculated using Eq. (3) (Methods 7.1) in different variants of the communication network are presented in Table 3. We present also the statistical significance for each measurement calculated using the *jackknife method* as in [50]. Below we interpret the cases when our network exhibits assortativity:

- `undir mutual` $r = 0.414$, **highly assortative** (similar result reported in [9]): the more reciprocal contacts you have, the more reciprocal contacts they themselves tend to have.
- `undir mutual` $r^w = 0.474$, **highly assortative**: the stronger reciprocal contacts you have, the stronger reciprocal contacts they themselves tend to have.
- Other coefficients were too low displaying no assortativity so we do not interpret them.

In summary, reciprocal communication is characterized by strong social status homophily where the status is defined by the number and the strength of contacts. However, this does not hold for unilateral communication where we notice even slight levels of heterophily, since `undir all` coefficients have small negative values.

Table 3
Degree assortativity coefficients in the communication network. Statistical significance is evaluated through standard deviation s calculated using the *jackknife method* [50]. The only result that is not statistically significant is in italics font.

Undirected network		r	s	Directed network	$r_{a,b}$	s
Unweighted	`undir mutual`	0.414	0.010	**in–in**	*−0.001*	*0.002*
				in–out	0.110	0.013
	`undir all`	−0.015	0.001	**out–in**	0.038	0.003
				out–out	0.389	0.014
		r^w	s		$r^w_{a,b}$	s
Weighted	`undir mutual`	0.474	0.017	**in–in**	−0.015	0.002
				in–out	0.207	0.020
	`undir all`	−0.014	0.001	**out–in**	0.014	0.004
				out–out	0.338	0.026

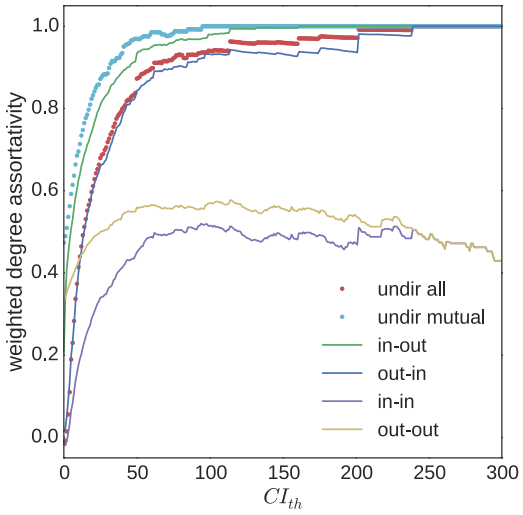

Fig. 3. *Degree assortativity* in the ensemble of thresholded `communication networks` as a function of minimum link strength given CI_{th}. `undir all` is r^w in undirected network including all links; `undir mutual` is r^w in undirected network with only reciprocal links; *in(out)–in(out)* are the four coefficients in directed networks $r^w_{in,in}$, $r^w_{in,out}$, $r^w_{out,in}$ and $r^w_{out,out}$.

4.2.2. Directed network variants

Directed networks allow to define four types of degree assortativity considering all the combinations of indegree and outdegree of the source and receiver nodes. The four assortativity coefficients introduced in Eq. (4) (Methods 7.1) show if the respective degree of a source node is correlated with the degree of the target nodes. In sociological terms, we are tackling relational analysis between source and receiver in communication [62]. As shown in Table 3, $r_{in,\,in}$ is the only negative of the four coefficients in our network. This is in agreement with the findings in directed followers Twitter network [49], except for $r_{out,\,in}$ which is also found negative in the followers network and it is slightly positive in our case. It is relevant mentioning that Myers et al. [49] calculated these coefficients in the followers graph as a correlation between logarithms of degrees, since otherwise they find almost no correlations. However, in our case we stay with the traditional definition, as in our network we do find strong correlations for some of the coefficients even without taking the logarithm of degree. The authors in [49] argue that Twitter exhibits negative assortativity coefficients, unlike other social networks, because of its role as an information network, too. Below we interpret the results in our network.

- $r_{in,out}$, $r^w_{in,out}$, **moderately assortative**: the more popular you are, the more active those who you contact (both in terms of number and strength of contacts).
- $r_{out,out}$, $r^w_{out,out}$, **highly assortative**: the more active you are, the more active the users whom you contact (the stronger your communication contacts, the stronger contacts those who you contact tend to have).
- Other coefficients were too low displaying no assortativity so we do not interpret them.

4.2.3. Assortativity as a function of link strength

We can create an ensemble of weighted communication networks by thresholding the original network on different minimum link strength. Then we calculate the above presented coefficients in each thresholded network. The result is degree assortativity as a function of minimum link strength based on CI_{th}, as shown in Fig. 3.

First insight is that already with a small increase in the threshold, the two assortativity coefficients that start slightly negative in the un-thresholded case (`undir all` and r_{in-in}) become positive. In other words, the network including weak links is not assortative on these two types of measures. However, as the weaker links are removed and stronger preserved, the assortativity sharply rises. With the threshold above 30 mentions, we detect high assortativity on all the coefficients. Another insight is that the pairs `undir all`, $r^w_{out,in}$ and `undir mutual`, $r^w_{in,out}$ follow similar patterns to each other. One possible interpretation is as follows. The pattern of active users contacting the popular ones described by $r^w_{out,in}$ is more of a unilateral pattern (due to similarity with `undir all`). On the other side, $r^w_{in,out}$ describes the pattern of popular users contacting active ones and seems to pertain more to reciprocal communication (due to similarity with `undir mutual`).

Moreover, looking at the higher CI_{th}, we notice two more interesting patterns. Two directed assortativity coefficients $r^w_{in,in}$ and $r^w_{out,out}$ start to slowly decrease, while the other four coefficients asymptotically reach maximum value 1. In our network case, the threshold of 239 mentions is when the four coefficients all become equal by reaching perfect assortativity and also when $r^w_{in,in}$ and $r^w_{out,out}$ become equal at value 0.505. Perfect assortativity reached in this way means that the network becomes fragmented in complete sub-networks, i.e., we will be left with small groups of users, possibly also pairs, that are all connected to each other by strong links. The fact that $r^w_{in,in}$ ($r^w_{out,out}$) never reach 1 means that we never arrive to the situation that all the users of equal popularity (activity) communicate. While not shown in Fig. 3, we calculated and those two coefficients continue to drop, while the others stay at the maximum value as we increase the threshold further. Piraveenan [58] suggested that directed networks can be classified using the combinations of their degree assortativity coefficient values. It is an interesting question for future work to analyze how here presented values in Twitter network compare to other social networks.

Overall analysis presented here also displays how structurally important changes in the Twitter network happen depending on the strength of links considered (from no assortativity to perfect assortativity as we move towards strong links). Bliss et al. [9] demonstrated temporal stability of degree assortativity in mutual mention network, while herein we demonstrate its variability with respect to the minimum link strength. This analysis also reaffirms the importance of analyzing interaction links compared to only binary links (such as friendship and followership) as is the case in most previous research. Coming back to the above mentioned negative assortativity results in the Twitter followers network [49], *we argue that at higher communication intensity and stronger links (requiring more time and effort than other interactions, such as following) the Twitter mention network serves more of a social than information role*. That is exhibited by the strong degree assortativity coefficients.

4.2.4. Burt's index

Another way to asses the social capital of nodes is through network constraint or Burt's index [13] (for the formula, see Methods 7.2). The distribution of Burt's index score is show in Fig. 4 (top left). A higher score means that the user has fewer contacts or that her contacts are densely interconnected hence limiting her view of the network for novel information. On the contrary, a lower score is given to the users near structural holes in the network who are exposed to diverse information and influences. That means that lower scores of Burt's index correspond to a higher social capital. Most of the users have score near 1, with a smaller peak near value 0.5 and a minority of rich users having the score close to zero. Hence, a majority of Twitter users in our mention network are confined within their local social circles and they are not exposed to alternative and novel information. Interestingly, the assortativity of mutual communication network on the Burt's

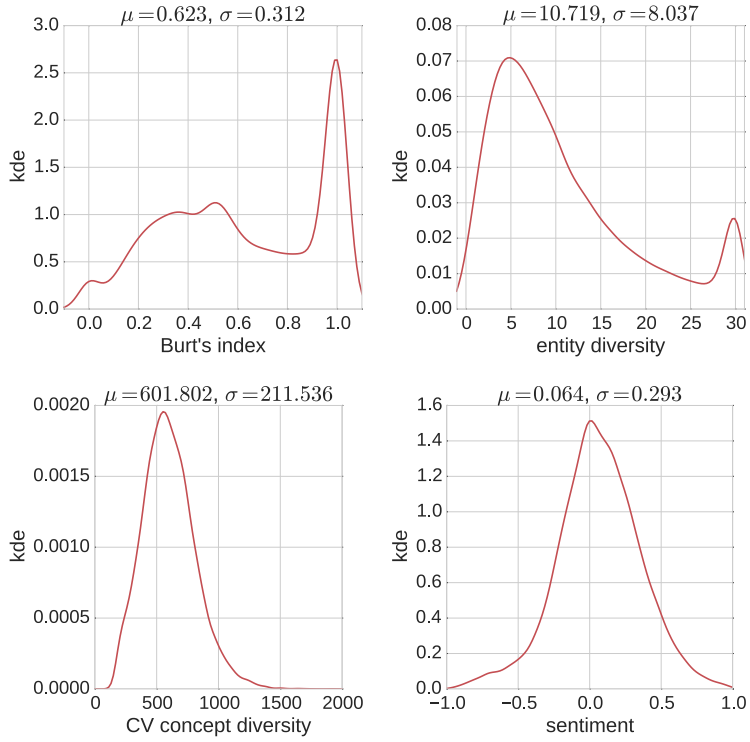

Fig. 4. *Social and semantic capital distributions*: kernel density estimates [55,63] of Burt's index, entity diversity, CV concept diversity and sentiment scores.

index is 0.67, so higher than the degree assortativity coefficients previously calculated. This result means that the nodes of, not only high degrees, but also of advantageous network position [13], preferentially connect. Such preferential attachment additionally increases the gap in the communication network between the information-poor and information-rich users.

To sum up, presented assortativity results reveal presence of *social status homophily* in communication (users with high status tend to assortatively connect) on different forms of social capital (number of contacts, activity, popularity, Burt's index). We also find slight amounts of *social status heterophily* in relation to unilateral links and popularity, but this heterophily quickly gives place to robust homophily when we shift the focus on the strong links.

4.3. Semantic status homophily

In addition to social capital, we can also define several forms of semantic capital in the communication network. Namely, the amount of diversity of user tweet contents with respect to a semantic attribute defines it semantic capital, similarly as in [65]. Hence we measure entity and CV concepts diversity (see Section 3.2.1 for the definition and Methods for detailed description of CVs). Another semantic feature that we introduce in capital assessment is sentiment score. In Fig. 4 we show kernel density estimates [55,63] of semantic capital distributions displaying heterogeneity of their values among users. While most of the users tend to have around 5 entities relevant to their tweet contents, we also find an important percent of users with nearly 30 such entities. Similarly for concepts, a majority of users has 500–700 concepts describing their tweets, but we find also users with with 1500–2000 concepts. As for sentiment, a majority of users tend to have neutral tweets sentiment, however, we also find users on both sides of the spectrum (negative and positive extremes). Hence, we conclude that there is large **semantic capital heterogeneity** among Twitter users (see [65] for similar result in other types of knowledge networks).

In Table 4, we show attribute assortativity on several semantic attributes. Concepts and taxonomy diversity refer to concepts and taxonomy levels retrieved using Alchemy API. For the topics, we also utilize the output from Alchemy API, which assigns a topic score relevance to each user's tweets. The assortativity results in the table suggest the presence of both, **value**, and **status** homophily. Since the assortativity is found in the unweighted network, this means that the semantic features are homophilous traits even for the weak links. Moreover, homophily holds for reciprocal

Table 4
Semantic status and value homophily: attribute assortativity r calculated using Eq. (2) (Methods 7.1) in unweighted communication network. Statistical significance is evaluated through standard deviation s calculated using the *jackknife method* [50] and shows that all the results are statistically significant.

Level	Status homophily				Value homophily			
attr	Wiki CVs diversity	Taxonomy diversity	Entity diversity	Concept diversity	Sentiment score	Topic music	Topic movies	Topic sex
Directed network, all edges								
r	0.144	0.157	0.292	0.173	0.315	0.151	0.136	0.136
s	0.003	0.003	0.003	0.003	0.003	0.003	0.004	0.004
Undirected network, mutual edges								
r	0.269	0.282	0.398	0.289	0.452	0.269	0.244	0.253
s	0.006	0.005	0.005	0.005	0.005	0.006	0.005	0.006

(mutual), and also unilateral (all edges) links. Once again, as with the degree assortativity, we find that mutual communication network exhibits notably higher levels of assortativity on all the analyzed attributes compared to the network including also unilateral links.

As the observed correlation levels could be induced by existing degree assortativity, we also test the presence of assortativity after randomization of the node's semantic attributes. The assortativity value in such case is importantly lower, 0.07 and so we conclude that indeed the communication network exhibits low to moderate levels of *semantic status and value homophily*. Moreover, among analyzed semantic attributes, status homophily is the largest with respect to entity diversity and value homophily with respect to sentiment.

4.4. Interplay between social and semantic capital

After confirming status homophily in terms of social and semantic capital, we investigate the relationship between these two forms of capital in the communication network. Whether the users who are rich in terms of social capital (and hence more network central) are also rich in terms of semantic capital (their tweets exhibit diversity on semantic aspects)? With this analysis, we respond to the call by authors in [65] to look for similar types of patterns as they have investigated in the bloggers and scientists networks. Testing for different forms of social against forms of semantic capital reveals no significant or low to medium correlations between them, as reported in [65]. In the end we conclude that the observed patterns in the Twitter communication network resemble more the bloggers than the scientists network presented in [65].

For the purpose of this analysis, we divide the social capital in terms of degree to popularity (indegree) and activity (outdegree). We visualize in Fig. 5 the joint distributions of popularity, activity and Burt's index, on one side, and the sentiment score and the entity and concepts diversity, on the other side. A wide spectrum of **entity diversity** is present for the users with, both, low and high popularity. Most popular users tend to be slightly more likely to have high semantic diversity. In the case of activity, we observe similar patterns that are a bit more pronounced. Basically, most actively communicating users are likely to have high entity diversity. However, we still find a number of users with diverse tweet contents that are not active. The highest correlation between

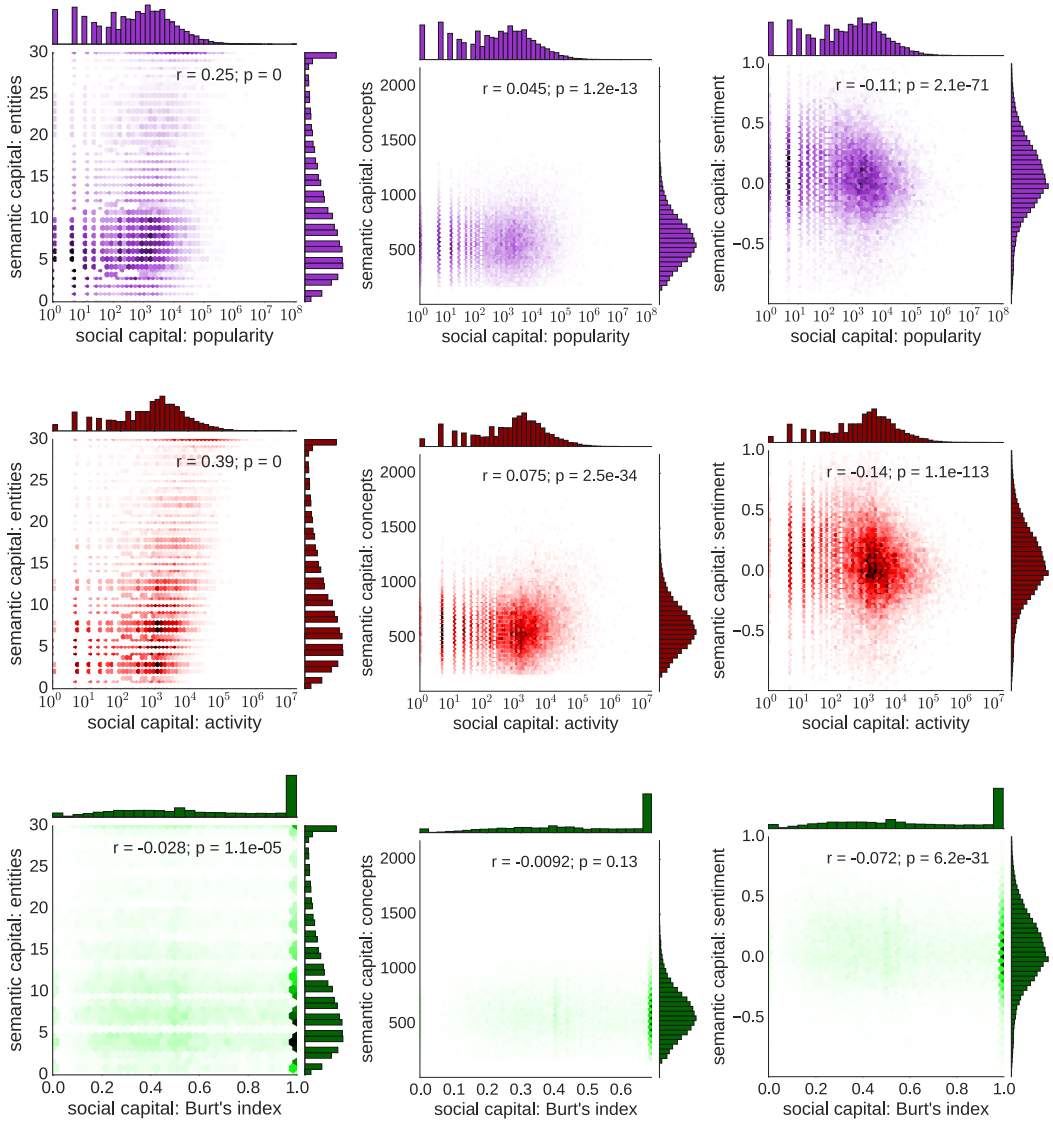

Fig. 5. *Joint distributions of social and semantic capital*: the darkness of the hexagon corresponds to the number of users with the combination of respective social and semantic capital values.

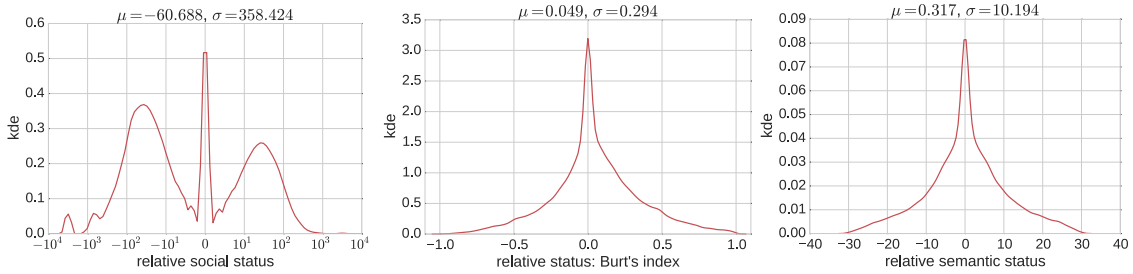

Fig. 6. Status differences in communication: Kernel density estimates [55,63] for distributions of popularity, Burt's index and semantic capital difference.

a form of social and semantic capital in our data is that between entity diversity and activity ($r = 0.394$). In the case of **CV concepts diversity**, we do not find any differences between popular and active users. The richest users in terms of both types of social capital tend to have an average semantic capital (between 500 and 1000). Burt's index does not correlate with any form of semantic capital. Hence, we conclude that different forms of semantic capital have different patterns of interplay with social capitals. The tweet **sentiment** of popular users is most likely to be neutral. On the other hand, the users who are towards positive or negative extremes of sentiment spectrum are likely to have low to modest popularity. Sentiment has slight *negative correlations* with both popularity and activity.

In summary, we find a form of semantic capital (concept diversity) that does not correlate with any form of social capital; and, a form of social capital (Burt's index) that does not correlate with any form of semantic capital. Semantic capital in terms of entity diversity has low to moderate positive correlation with popularity, and the correlation increases with activity. Sentiment correlates negatively with both, popularity, and activity.

4.5. Relative status of source and receiver

Relative status is an additional way to investigate social and semantic capitals. Relative status is simply the difference in status of the source and receiver on a particular form of capital. In Fig. 6 we show the distributions of relative social status (popularity and Burt's index) and relative semantic status (entity diversity) between source and receiver in communication.

Both distributions for relative social status exhibit a dominant peak at zero (users with similar status are most likely to communicate). However, plotting the popularity difference on a log scale reveals two additional interesting peaks at intervals $(-100, -10)$ and (10,100). There is a higher likelihood for users with differences in popularity belonging to these ranges to be talking to each other. The left peak is higher, and this together with the negative mean value informs that source users tend to be a bit less popular. There is also a small number of users mentioning considerably more popular users than themselves (leftmost part of the distribution). This happens to a smaller extent in the other direction, from more popular source users. When it comes to Burt's index, the mean is slightly positive. However, since higher values of Burt's index represent lower status, this means that again users of lower social status are slightly more likely to initiate communication.

The general pattern is similar for semantic status: most of communication happens between those who have close to equal semantic status. A positive mean in this case, however, shows it is slightly more likely for semantically rich users to initiate communication.

For the joint distribution of social and semantic relative statuses we find (analyzed, not shown) a wide range of combinations. There is a small positive correlation between the two. As for the small number of users who initiate communication towards a considerably more popular users discussed above, we find that they tend to be semantically richer compared to the receiving users. We speculate that this *semantic superiority might be a needed approach for such users to compensate for their lower popularity*.

4.6. Status inconsistency of source and receiver

Finally, we can tackle a sociological proposition that source and/or receiver **status inconsistency** can increase effectiveness of their communication [62]. *Status inconsistency (internal heterophily of an individual) is defined in sociology as the relative lack of similarity in an individual's ranking on various indicators of social status* [39]. This is suggested to be an attribute of individuals who are drivers of social change [39]. Hence we introduce status inconsistency for Twitter users as a relative difference in their social and semantic capital ranks. We apply a similar formula to calculate **user status inconsistency** (st_{inc}) as in [39]:

$$st_{inc} = \begin{cases} -(1 - r_{soc}/r_{sem}), & \text{if } r_{soc} \le r_{sem} \\ (1 - r_{sem}/r_{soc}), & \text{otherwise;} \end{cases}$$

where r_{soc} and r_{sem} are users ranks in terms of social and semantic capital, respectively, among all users. This definition allows firstly to asses the amount of user status inconsistency (how close is $abs(st_{inc})$ to 1), and second, it also encodes whether she has higher social (st_{inc} positive) or semantic (st_{inc} negative) status.

While we can not measure effectiveness of communication directly using our dataset, we allow *communication intensity* to be a proxy for it. Our hypothesis in this regard is: the higher the communication intensity between a source and receiver, the higher potential for an effective communication.

For directed links $(e_{i,j})$ we define **link status inconsistency**:

$$st_{inc}(e_{i,j}) = st_{inc}(u_i) \cdot st_{inc}(u_j),$$

where $st_{inc}(u_i)$ and $st_{inc}(u_j)$ are user status inconsistency of the source and receiver, respectively.

This simple formula produces a higher absolute value for the links with higher total inconsistency. The sign in this case indicates whether the source and receiver are ranked highly on the same form of capital ($st_{inc}(e_{i,j})$ positive) or on different forms of capital ($st_{inc}(e_{i,j})$ negative).

We indeed find a significant correlation between the introduced link inconsistency and communication intensity ($r = 0.27$, $p < 0.01$). Fig. 7 indicates the following: the communication between two users tends to increase with status inconsistency of one or both of the users, if they are both rich on the same form of capital. If the users are status inconsistent but rich on different forms of capital, then their communication intensity tends to decrease. As with other findings regarding capitals, the patterns are pronounced for the extreme case (strong links) while there is a wide spectrum of link inconsistency values taken by the rest of the links (Fig. 7, left).

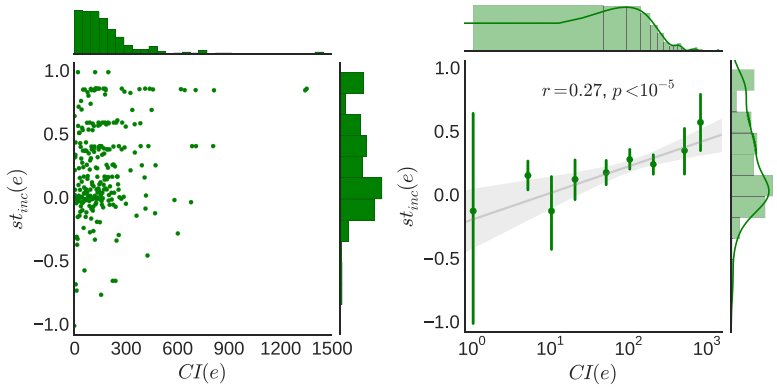

Fig. 7. Relationship between communication intensity and link inconsistency: (left) scatter plot; (right) linear regression visualization – we apply logarithmic binning to account for long-tailed distribution of $CI(e)$; average value and standard deviation are shown for each bin.

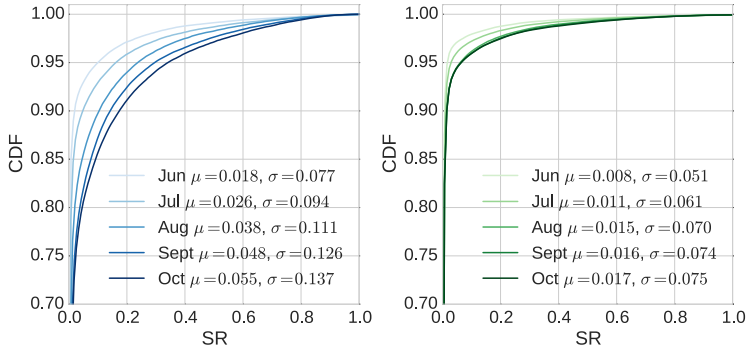

Fig. 8. Cumulative SR distributions for 5 months: (left) in mutual `communication network` and (right) in the rest of `SR network`, i.e., among not connected pairs. For better visualization of the differences in distributions the y-axis is thresholded above 0.7. The distributions are together sharply rising up to around that point. .

5. Temporal evolution of semantic homophily

In previous sections we performed analysis on a snapshot of Twitter network formed from the whole 6 months dataset. In this section we investigate temporal aspects of semantic homophily by looking at different snapshots of the network for each month. First we investigate temporal change of the SR values. Cumulative distribution functions (CDF) of SR values for each full month in our dataset are shown in Fig. 8 for the mutual communication network and for the rest of the links in the full SR network. Precisely, for the distribution in left plot we consider only the reciprocal links, while for the right plot, we take the difference between the full SR network and the full communication network. In this way, we try to separate user pairs involved in bilateral communication (possibly affected by social influence) and those between whom no communication of any type occurred (less likely to be affected). Gradual increase in SR values takes place in both cases over time. For example, for mutual communication for June, CDF reaches 0.95 near SR value 0.1. This means that only 5% of user pairs have higher SR than 0.1 during June. The corresponding distribution for October has CDF value 0.85 around the same SR of 0.1. Hence, in October, around 15% of user pairs have higher similarity than 0.1. To confirm that this is not only an effect of visualization, but that the distributions indeed change, we apply Kolmogorov–Smirnov (K–S) test [43]. K–S test applied on two empirical distributions informs about the hypothesis that they are drawn from the same theoretical distribution. Again comparing the distributions for June and for October, K–S statistics gives $p < e^{-24}$ for the mutual communication network and $p < e^{-197}$ for the rest of the SR network. Hence K-S test in both cases strongly rejects the hypothesis that the distributions are drawn from the same.

The increase in average SR among not interacting users (Fig. 8, right) is peculiar. It indicates a possible external influence taking place during the period causing all users to talk more on a similar (external) topic. We investigated semantic layers and found several commonly discussed topics among our users (one popular series aired at that time, several music bands and couple of real world events). However, since Twitter that we investigate is not the only possible medium for our users to communicate and influence each other, this does not allow us to assert whether the increase is indeed (only) due to external influence. This finding is similar to the reported increase in sharing information which user's friends have shared, even when the user is not exposed to that information from her friends [7].

The average increase of SR in the mutual communication network (Fig. 8, left) can be due to homophily, i.e., new user pairs starting communication. As we presented in Section 4.1, communication propensity is larger for user pairs with high SR. Hence, if pairs who are more similar are preferentially getting connected, then this could explain the increase in SR among connected users. Another explanation could be social influence, i.e., that already connected pairs are becoming more related. Sociology also suggests to look for link dissolution among dissimilar individuals [10,24] as another explanation.

We continue by investigating formation and dissolution of links through time. The requirement for active engagement from both source and receiver allows us to define communication activation (link formation) and communication decommission (link dissolution) for reciprocal links. For each of the 69,312 reciprocal links observed during the whole period, we define **communication activation (formation)** time to be the month when for the first time both users have mentioned each other (in our dataset period). **Communication decommission (dissolution)** time is given by the last month in our dataset that the users have both mentioned each

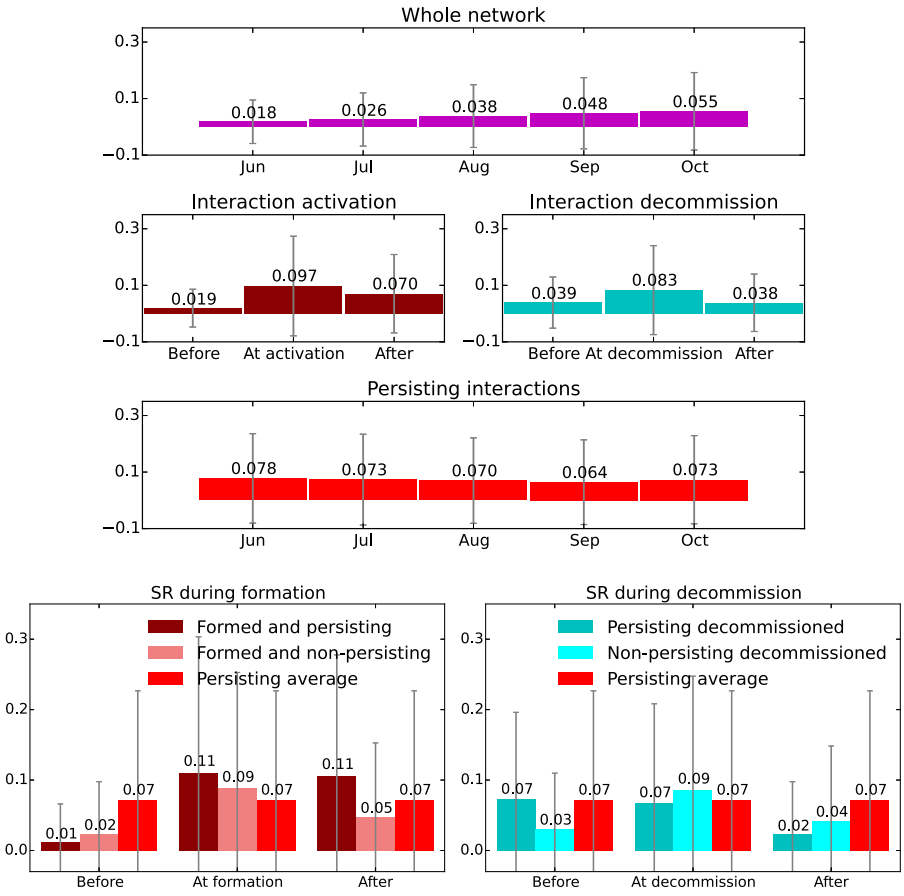

Fig. 9. SR evolution on the links. We show average SR for the whole network, for the formed, decommissioned and persisting links. Then we also separate formed links to those that persist and those that do not; and decommissioned links to those that were previously persisting or not. For comparison, average SR of persisting links is also shown. Error bars show standard deviation values.

other, after which one or both sides ceased the communication. In order to have enough data to calculate users similarity prior/after to links activation/decommission, we require the month of activation/decommission to be between July and September. With this approach, we find in total 13,492 link activations and 10,080 link decommissions in our dataset. As a first insight, we notice that slightly more links are activated than decommissioned.

5.1. SR and evolution of links

Temporal change of average SR on links prior to and after the *formation* is shown in Fig. 9. The SR between a user pair noticeably increases at the month of their communication activation. Similar result has been found in other networks, for instance among Wikipedia admins [17] and for Flickr users [78]. The drop in average SR in the period after the link activation is also reported in earlier studies [78]. To investigate the drop in our case, we look for an evidence that some interactions might not be preserved for long. This is one aspect where our approach is advantageous compared to the previous approaches that consider a one time link formation (adding someone as a friend or following) and do not require an active engagement afterwards.

Indeed, we find in total 8166 links that are *activated and then also decommissioned* during our dataset period. The SR change separated between such links and those that persist after the formation is shown in the plot titled *SR during formation* in Fig. 9. The average SR for formed and persisting links stay high after they are formed. It is those links that will get decommissioned soon that contribute to lowering the total average SR after the formation. This result displays that homophily needs to be considered together with active engagement, link strength and their temporal dynamics.

If observing only the *persisting links* (also Fig. 9) that were already active and continued during the whole period, we notice a relatively stable average SR through time despite that the average SR in the whole network has increased from June until October. Also, SR on persisting links is higher compared to the whole network. The stability of SR for an established communication could suggest a lack of influence in our network. However, we are careful with such an interpretation, since this result might also indicate a saturation effect taking place. If looking at newly formed links which persist and have high SR (0.11), that indicates how at first, the users might influence each other for some time. However, their similarity is likely to stabilize around this specific SR value (~ 0.07) for persisting links in our dataset.

Fig. 9 also displays temporal change of SR on links that get *decommissioned*. Here, too, we separate the links that get decommissioned to those that were previously *persisting* from those that have formed during our dataset time frame, i.e., *non persisting*. Indeed, we notice how the persisting links have the above mentioned characteristic average SR of 0.07 which does not change during the actual month of decommission, but afterwards drops significantly to ~ 0.02. The non-persisting links reach even higher SR during the month of decommission, but before and after their SR is lower. This can indicate a sort of short-lived active engagement/interest between such pairs, unlike a more stable relationship on previously persisting links. The drop in average SR on the links that get decommissioned is striking: SR becomes from 2 (on non-persisting) to 3 (on persisting) times lower after link dissolution. However, since the SR values on previously persisting links

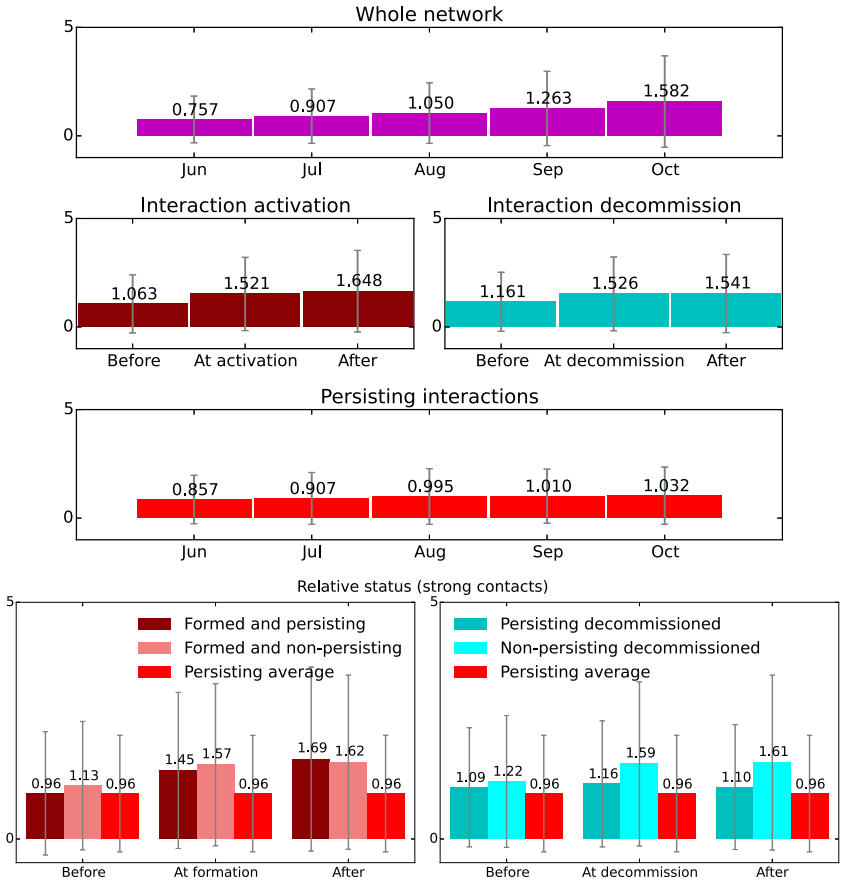

Fig. 10. Relative status evolution on the links (difference in number of reciprocal contacts). We show average relative status for the whole network, the formed, decommissioned and persisting links. Then we also separate formed links to those that persist and those that do not; and decommissioned links to those that were previously persisting or not. For comparison, average relative status of persisting links is also shown. Error bars show standard deviation values.

before their decommission are around the same as on the links that stay persisting, then *in terms of SR there is no observable dissimilarity between users before they will cease communication*. This is in contrast to the expectation from sociology [24]. We investigate other plausible reasons for the link dissolution below.

5.2. Relative social status and evolution of links

Operating on the same sets of links as so far, we now look at **social capital** of the communicating user pairs. As presented earlier, different forms of social capital can be assessed. Since herein we look at reciprocal communication, it is natural to asses social capital in terms of the number of *reciprocal contacts*. In Fig. 10, we show the evolution of the relative status between communicating user pairs.

Relative social status (discussed in Section 4.4) is defined as the absolute *difference between social capitals* of source and receiver users. In Fig. 10, we first notice the difference between persisting links and other types of links and also between the whole network average. Persisting links have lower relative status, i.e., users who are actively communicating tend to have similar social status ranks. While homophily on the status level is not new and we have discussed it (in Section 4.5), herein we extend the understanding of it. Namely, we exhibit its underlying mechanisms in communication network: both types of links, those that are newly formed and those that will get decommissioned, have slightly, but notably higher relative social status compared to persisting links. Hence, we find evidence that *link dissolution happens due to dissimilarity in social status*. Another interesting observation is that user pairs that start with higher relative status compared to persisting also get decommissioned later (while those who start around that persisting average indeed persist communication later).

To reiterate, our analysis on SR and social status during evolution of links so far gives two insights about the decommissioned links:

- no semantic differences (the SR is not lower at the time when link dissolution happens compared to the persisting average), and
- higher status differences (also compared to persisting links).

Hence, there is indication in the Twitter network that *persisting communication links dissolve in the presence of status level heterophily rather than value level heterophily*.

In conclusion, presented types of interactions show the importance of considering both homophily and influence as dynamic interdependent tendencies [77] in temporal networks, instead of looking at static snapshots. Our analysis on interaction decommission reveals similar results as in [53] where it is showed how not accounting for homophily effect on link dissolution may importantly affect social influence estimation. Precisely, we suggest that on a same communication link (interaction) at different points of time with reference to its activation/decommission time, one or the other of the tendencies might be playing a stronger role. Our dataset time-frame does not allow for that, but for future work, we aim to look at the period in which link formations and deletions might be happening, and whether there are some natural cycles in the human communication network.

6. Discussion and conclusion

The research on homophily is vast and growing, especially so nowadays with the availability of OSN data. Nevertheless, the mechanisms that interweave people in social networks are not yet fully explored and understood. We explore this question in a particular type of social network – communication network, and by focusing on a specific homophilous trait – semantics. However, the framework we present can be easily adapted, extended and applied on other type of social networks and for different types of homophily.

Our first set of findings quantifies to what extent semantic homophily and social influence affect the propensity and intensity of communication, though we are not trying to distinguish between these two factors. Taking interaction frequency as a proxy for link strength, we show the differences in homophily between strong and weak links. Concretely, we exhibit that semantic relatedness is a robust homophilous trait for the strong links and less so for the weak links.

By testing for existence of status and value homophily, we confirm that these general theories from sociology hold in a communication social network. We detect value homophily for tweet topics and sentiment, and status homophily for social and semantic capital of users. We also investigate the relationship between the two types of capitals. While popularity and semantic capital are positively correlated, sentiment, inversely, is negatively correlated with social capital. Burt's index seems to be completely unrelated with semantic aspects of communication. In any case, we exhibit large diversity among users on the existing combinations of capitals they posses. Additional investigation on sociological concept of relative status reveals strong preference for communication with users of similar status. However, for relative social status, popularity in particular, we notice a pattern of less popular users initiating more communication towards higher popularity users. We also find evidence for sociological proposition that status inconsistency of one or both of the parties increases communication effectiveness. To recall, status inconsistent users rank highly on one dimension of social status and lowly on another dimension. The dimensions we investigated are popularity and semantic status. Moreover, our data suggest additional hypothesis: this proposition holds only when both users rank highly on the same dimension of social status, otherwise, communication intensity and link strength decrease.

Temporal analysis of communication reveals that persisting links feature remarkably stable properties, such as semantic and status similarity between source and receiver and this is not the case for short-lived links. We also find evidence that the tendencies of homophily and influence are dynamic and change their role and magnitude in time. Besides confirming previous findings in other types of networks that similarity between users sharply grows before their link formation, we also explain in part the following decrease in similarity – as a result of later decommission of some of the links. A novel insight we make is that relative difference in social status is a stronger predictor for link decommission compared to differences on a value homophily level. While the results we presented with respect to different types of links show robustness in our dataset, further studies are needed that will evaluate similar questions on datasets with longer time-scales to confirm these findings.

Overall, our analysis on Twitter contributes to understanding the role of semantic value, and social and semantic status homophily in online communication. We hope that it will inspire future research in other OSNs towards building a more comprehensive picture on the social tendencies that shape online communication (Table 5).

6.1. Limitations

Several limitations are posed to our work because of the dataset used. First, the results are informing only about computer-mediated communication through Twitter and we do not tackle the impact of the online medium on social interaction. Second, there is a special nature to the Twitter communication and the mention network represents only one of the channels used within the OSN. Hence another limitation is posed by the restricted dataset as we are not considering the entire Twitter channel for information flow. There is also a considerable amount of information flowing along the retweet network, which is not taken into consideration in this work. We refer to the analyzed mention network as communication network, however, we also acknowledge that some mentions are used to expose communication to a large public, hence not representing bilateral communication. Besides this, the mention mechanism in Twitter can be sometimes biased towards specific target audiences for specific information [71]. Moreover, our temporal analysis on the links is made for a limited period of 6 months. While different types of links do exhibit different properties with respect to homophily, we acknowledge that the robustness of presented findings should be evaluated on a longer time-scale. However, our results in this specific OSN context are still relevant as they provide grounds for comparison to similar analyses in some other OSNs. Only a set of studies performed with data from diverse settings can inform us more about homophily in online communication and eventually in general human communication.

Table 5
Summary of our contributions: for each theory or question from sociology that defined the analysis we describe found evidence and/or some hypotheses or open questions that arise from the analysis.

Theories from sociology	Experimental evidence	Hypotheses
Quantification		
Homophily [44] **SWT theory** [28]	Semantic homophily larger for strong links.	
Status level homophily [35] **Structural holes** [13]	Degree assortativity increases with link strength. Network assortative on Burt's index.	
Value level homophily [35]	Assortativity on forms of semantic captial.	
Social capital [59]	Popularity and activity increase with semantic captial. Popularity and activity correlate negatively with sentiment.	Burt's index does not correlate with semantic capital.
Status inconsistency [39] **SWT theory** [28]	Status inconsistent links are stronger when both users are high on the same status dimension;	otherwise status inconsistent links are weaker.
Temporal evolution		
Homophily [44]	Average user similarity in network increases in time.	Average user similarity even of those not connected increases in time.
Status/value level homophily[35]	Users similar on semantic value level likely to connect.	Similarity of persisting links on status/value level is stable through time.
Dissolution of heterophilous links [24]	Confirmed;	but only for heterophily on status level.

7. Methods

7.1. Assortativity formulas

Assortativity [50] is defined slightly differently depending on the type of node property considered. If the property is **discrete** (it defines classes of nodes), then

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 + \sum_i a_i b_i}, \tag{1}$$

where e_{ij} represents a fraction of edges in the network that connect a node with property i to a node with property j, while $a_i = \sum_j e_{ij}$ and $b_j = \sum_i e_{ij}$. Note that $a_i = b_j$ when the network is undirected.

If the node property is **scalar**, such as semantic diversity, then assortativity is calculated as

$$r = \frac{\sum_{ij} ij(e_{ij} - a_i b_i)}{\sigma_a \sigma_b}, \tag{2}$$

where σ_a and σ_b are standard deviations of a_i and b_j.

Assortativity values range from 1 in a perfectly assortative network to −1 in a perfectly dissasortative network. Any discrete or scalar attribute of nodes can be used to calculate this coefficient.

Undirected degree assortativity r is an inherent second order metric [54] of any network. In the formula introduced by Newman [50], the exact property considered in calculation is *remaining degree* q_k of node, which is one less than the actual degree p_k. The reason is that when assessing all possible pairs of nodes in the calculation, we are interested in how many other links they have than the one between them. Degree assortativity in an undirected network is given by the formula below:

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{{\sigma_q}^2}, \tag{3}$$

where σ_q is standard deviation of q_k and $\sum_j e_{jk} = q_k$.

For directed networks, Piraveenan et al. [57] introduce additional formulation to the one given by Newman [50], allowing for four types of **directed degree assortativity**: $r_{in,\,in}$, $r_{in,\,out}$, $r_{out,\,in}$ and $r_{out,\,out}$. We use the formulas introduced by them and also later applied on a Twitter followers network by Myers et al. [49], as given below:

$$r_{a,b} = \frac{\sum_{jk} jk(e^{ab}_{jk} - q^a_j q^b_k)}{{\sigma_q}^a \sigma^b_q}, \quad \text{for} \quad a, b \in \{in, out\}, \tag{4}$$

where q^{in}_j and q^{out}_j are, respectively remaining indegree and outdegree of a source node and q^{in}_k and q^{out}_k are, respectively remaining indegree and outdegree of a target node.

Finally, one can consider weights for **weighted degree assortativity** r_w calculation, as introduced in [40]:

$$r^w = \frac{H^{-1}\sum_\phi (w_\phi \prod_{i\in F(\phi)} q_i) - ((H^{-1}/2)\sum_\phi (w_\phi \prod_{i\in F(\phi)} q_i))^2}{(H^{-1}/2)\sum_\phi (w_\phi \prod_{i\in F(\phi)} q_i^2) - ((H^{-1}/2)\sum_\phi (w_\phi \prod_{i\in F(\phi)} q_i))^2}, \tag{5}$$

where w_ϕ denotes the weight of the ϕth link, $F(\phi)$ is the pair of nodes connected by the ϕth link, and H is the total weight of all links in the network. If all weights in the network are equal then r_w reduces to that of an undirected network, r.

7.2. Burt's network constraint index

For a given node i, its network constraint index is given using formula:

$$C = \sum_j c_{ij}, \tag{6}$$

where c_{ij} is the extent to which i's network is directly or indirectly invested in its relationship with node j. This extent is calculated using formula:

$$c_{ij} = (p_{ij} + \sum_q p_{iq} p_{qj})^2, \tag{7}$$

for $q \neq i, j$, where p_{ij} is the portion of i's time and energy invested in contact j. This value is calculated as:

$$p_{ij} = z_{ij} / \sum_q z_{iq}, \tag{8}$$

where z_{ij} measures normalized connection strength between i and j.

7.3. Semantic database

We build the semantic database according to existing Explicit Semantic Analysis (ESA) algorithm [25]. ESA uses Wikipedia as a source of semantic knowledge. In this section we describe how we build the semantic database using an English pages Wikipedia dump from April 2015 (52GB in size, uncompressed). The first step is to take the article texts as the algorithm builds on the large amount of knowledge they provide. We then apply an open-source script *wikiextractor* [26] to pre-process and clean the texts. The ESA algorithm is based on the term frequency - inverse document frequency (TF-IDF) [6] scores of words in different articles in the Wikipedia corpus. As a result a word w_1 is mapped to the *concept vector* $CV(w_1) = \{(C^1_1, V^1_1),\ (C^1_2, V^1_2),\ (C^1_3, V^1_3), \ldots,\ (C^1_{M1}, V^1_{M1})\}$. C^1_j represent Wikipedia concepts and V^1_j are TF-IDF scores for the word w_1 in those articles and they are calculated as follows:

$$V^1_j = TF \cdot IDF = (1 + \log(f_{1,j})) \cdot \log(N/n_t), \tag{9}$$

where TF is the log-normalized raw frequency ($f_{1,j}$) of the word w_1 in article j, and IDF is the inverse document frequency, N is the number of articles, and n_t is the number of articles in which the word w_1 is present.

The algorithm was implemented in Python with application of the scikit-learn machine learning library [56] and the resulting database was stored in a MongoDB collection. Since some of the concept vectors might have tens of thousands of terms; prior to storing, we apply the pruning process [25] that for each word keeps only important CV elements. The algorithm implementation needs tuning several parameters, and in this process we also consult some of the existing implementations of the ESA algorithm. Our implementation of ESA[6] is open-source and published on Github [67].

7.3.1. Word semantic relatedness

The SR between words is determined through a set of concepts highly related to them [25,31]. Let us assume that the SR between words w_1 and w_2 is requested. The $SR(w_1, w_2)$ calculation follows the two steps below.

- **Determining the corresponding CVs derived from Wikipedia for the words** w_1 and w_2. The CVs are based on concepts (or articles) from Wikipedia which are related to the words. Let us assume that w_1 is mapped to: $CV(w_1) = \{(C^1_1, V^1_1),\ (C^1_2, V^1_2),\ (C^1_3, V^1_3), \ldots,\ (C^1_{M1}, V^1_{M1})\}$ and w_2 is mapped to: $CV(w_2) = \{(C^2_1, V^2_1),\ (C^2_2, V^2_2),\ (C^2_3, V^2_3), \ldots,\ (C^2_{M2}, V^2_{M2})\}$. In the following, we will assume that N is the number of common concepts in $CV(w_1)$ and $CV(w_2)$.
- **Calculating cosine similarity between obtained CVs**. This measure gives the cosine of the angle between the two vectors

[6] https://github.com/sanja7s/SR_Wiki_ESA

Table 6
SR knowledge database evaluation.

Human judgments dataset	Spearman's rank	Pearson's correlation
WordSim-353	0.51	0.45
Miller and Charles	0.79	0.82
Word pair similarity, MTurk	0.53	0.45
Rubenstein and Goodenough	0.81	0.74
MEN dataset of word pair sim.	0.73	0.44
Average	**0.67**	**0.58**

$CV(w_1)$ and $CV(w_2)$. The cosine measure can be re-formulated for our purpose as follows:

$$SR(w_1, w_2) = \cos(CV(w_1), CV(w_2)) = \frac{\sum_{i=1}^{N} V_i^1 \cdot V_i^2}{\sqrt{\sum_{k=1}^{M1} \left(V_k^1\right)^2} \cdot \sqrt{\sum_{l=1}^{M2} \left(V_l^2\right)^2}}, \quad (10)$$

where i iterates over the common concepts.

The $SR(w_1, w_2)$ values range from 0 (i.e., no semantic relatedness) to 1 (i.e., perfect semantic relatedness).

7.3.2. Document semantic relatedness

The SR between documents is measured through the SR of the words found in the documents. Let us assume that the SR between documents d_1 and d_2 is requested. The $SR(d_1, d_2)$ calculation follows the three steps below.

- **Apply term frequency (TF) to find the frequency of words in the document**. The result of this step is a list of words with their corresponding TF scores. Let us assume the mapping: $d_1 \Rightarrow T(d_1) = \{(t_1^1, v_1^1), (t_2^1, v_2^1), (t_3^1, v_3^1), \ldots, (t_m^1, v_m^1)\}$, $d_2 \Rightarrow T(d_2) = \{(t_1^2, v_1^2), (t_2^2, v_2^2), (t_3^2, v_3^2), \ldots, (t_n^2, v_n^2)\}$, and $m < n$.
- **Determining the corresponding** *CVs* derived from Wikipedia for the documents d_1 and d_2**.** For each term in the lists $T(d_1)$ and $T(d_2)$ we derive their individual *CVs* (as described for words in Section 7.3.1). When summarizing the *CVs* for one document, the *CV* for each term is multiplied with its *TF* score in the document (found in the previous step). If the terms in $T(d_1)$ have the same concepts in their *CVs*, we sum the weighted TF-IDF scores of those concepts. After this process we obtain $CV(d_1)$, the list of Wikipedia concepts and TF-IDF scores which are related to all the terms in $T(d_1)$. Similarly, for d_2 we find $CV(d_2)$.
- **Calculating cosine similarity between obtained CVs**. Finally, we obtain the $SR(d_1, d_2)$ between documents by calculating the cosine similarity of $CV(d_1)$ and $CV(d_2)$ (see Eq. 10).

7.3.3. SR database evaluation

The English version of Wikipedia used includes over 2.5 million articles. Since many of the articles are highly specialized, and due to the described pruning process, we find only around 15% of those articles (387,992) relevant for our tweets corpus. In a similar manner as in the original paper [25], we evaluate the quality of the SR database that we built against available datasets with human judgment for word pairs relatedness. We use several such datasets available online, as one of the most comprehensive current resources [23]. The results of the evaluation are presented in Table 6. We do not provide herein a comparison with the existing implementations, since not all of them provide their evaluation on the same datasets with human judgments, and since a previous study comparing them has shown that some of these results are incompatible [16]. However, our evaluation scores are comparable to the original implementation [25] and to the ESA implementations available online.

Acknowledgments

S.Š. research was partially financed by CIVIS EU FP7 (608774) project (FP7- SMARTCITIES-2013). I.M. work was partially financed by the Faculty of Computer Science and Engineering at the University Ss. Cyril and Methodius. S.Š. and I.M. also gratefully acknowledge the CyberTrust research project for their support. B.G. thanks the Moore and Sloan Foundations for support as part of the Moore-Sloan Data Science Environment at New York University. P.H. thanks General Research Fund 26211515 from the Research Grants Council of Hong Kong. The work of TH. Nguyen was supported by the Academy of Finland, decision number 295694. S.Š. thanks A. Ukkonen for the help with the SR database implementation. S.Š. also acknowledges collaboration with P. T. Trung during his MSc thesis project when we performed a similar type of SR analysis on Twitter data. The authors also thank A. Gionis for the helpful discussion and for reviewing the manuscript.

References

[1] A. Abbasi, A longitudinal analysis of link formation on collaboration networks, J. Informetr. 10 (3) (2016) 685–692.
[2] An IBM Company, Alchemyapi and IBM Watson, 2016. http://www.alchemyapi.com/api.
[3] A. Anagnostopoulos, R. Kumar, M. Mahdian, Influence and correlation in social networks, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2008, pp. 7–15.
[4] S. Aral, L. Muchnik, A. Sundararajan, Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks, Proc. Natl. Acad. Sci. 106 (51) (2009) 21544–21549.
[5] S. Aral, D. Walker, Identifying influential and susceptible members of social networks, Science 337 (6092) (2012) 337–341.
[6] R. Baeza-Yates, B. Ribeiro-Neto, et al., Modern Information Retrieval, vol. 463, ACM Press, New York, 1999.
[7] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, The role of social networks in information diffusion, in: Proceedings of the 21st International Conference on World Wide Web, ACM, 2012, pp. 519–528.
[8] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, O'Reilly Media, Inc., 2009.
[9] C.A. Bliss, I.M. Kloumann, K.D. Harris, C.M. Danforth, P.S. Dodds, Twitter reciprocal reply networks exhibit assortativity with respect to happiness, J. Comput. Sci. 3 (5) (2012) 388–397.
[10] P. Block, T. Grund, Multidimensional homophily in friendship networks, Netw. Sci. 2 (02) (2014) 189–212.
[11] J. Bollen, B. Gonçalves, G. Ruan, H. Mao, Happiness is assortative in online social networks, Artif. Life 17 (3) (2011) 237–251.
[12] P. Bourdieu, The forms of capital (1986), in: Imre Szeman, Timothy Kaposy (Eds.), Cultural Theory: An Antholog, Wiley-Blackwell, 2011, pp. 81–93.
[13] R.S. Burt, Structural Holes: The Social Structure of Competition, Harvard University Press, 2009.
[14] N.A. Christakis, J.H. Fowler, Social contagion theory: examining dynamic social networks and human behavior, Stat. Med. 32 (4) (2013) 556–577.
[15] M. Conover, J. Ratkiewicz, M.R. Francisco, B. Gonçalves, F. Menczer, A. Flammini, Political polarization on twitter, Proceedings of International Conference on Web and Social Media, ICWSM, 133, 2011, pp. 89–96.
[16] I. Cramer, How well do semantic relatedness measures perform?: a meta-study, in: Proceedings of the 2008 Conference on Semantics in Text Processing, Association for Computational Linguistics, 2008, pp. 59–70.
[17] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, S. Suri, Feedback effects between similarity and social influence in online communities, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2008, pp. 160–168.
[18] S. Currarini, M.O. Jackson, P. Pin, An economic model of friendship: homophily, minorities, and segregation, Econometrica 77 (4) (2009) 1003–1045.
[19] S. Currarini, J. Matheson, F. Vega-Redondo, A simple model of homophily in social networks, Eur. Econ. Rev. 90 (2016) 18–39.
[20] M. De Choudhury, Tie formation on twitter: Homophily and structure of egocentric networks, in: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, IEEE, 2011, pp. 465–470.
[21] M. Dehghani, K. Johnson, J. Hoover, E. Sagi, J. Garten, N.J. Parmar, S. Vaisey, R. Iliev, J. Graham, Purity homophily in social networks, J. Exp. Psychol.: General 145 (3) (2016) 366.
[22] N.B. Ellison, et al., Social network sites: definition, history, and scholarship, J. Comput.-Mediat. Commun. 13 (1) (2007) 210–230.
[23] M. Faruqui, C. Dyer, Community evaluation and exchange of word vectors at wordvectors.org, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Baltimore, USA, 2014.
[24] D. Felmlee, S. Sprecher, E. Bassin, The dissolution of intimate relationships: a hazard model, Soc. Psychol. Q. (1990) 13–30.

[25] E. Gabrilovich, S. Markovitch, Wikipedia-based semantic interpretation for natural language processing, Journal of Artificial Intelligence Research 34 (2009) 443–498.
[26] Giuseppe Attardi, Wikipedia Extractor, 2015. https://github.com/attardi/wikiextractor.
[27] Karthik Gopalakrishnan, Arun Pandey, Joydeep Chandra, Social Interaction in Flickr: Mechanisms and Role in Link Formation, CoRR (2015). arxiv: http://arxiv.org/abs/1508.07577
[28] M.S. Granovetter, The strength of weak ties, Am. J. Sociol. 78 (6) (1973) 1360–1380.
[29] Y. Halberstam, B. Knight, Homophily, Group Size, and the Diffusion of Political Information in Social Networks: Evidence from Twitter, Technical Report, National Bureau of Economic Research, 2014.
[30] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, Semantic similarity from natural language and ontology analysis, Synthesis Lectures on Human Language Technologies 8 (1) (2015) 1–254 Morgan & Claypool Publishers.
[31] N.T. Hieu, M. Di Francesco, A. Ylä-Jääski, Extracting knowledge from wikipedia articles through distributed semantic analysis, in: Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies, ACM, 2013, p. 6.
[32] C.J. Hutto, E. Gilbert, Vader: a parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 2014.
[33] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media? in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 591–600.
[34] T. La Fond, J. Neville, Randomization tests for distinguishing social influence and homophily effects, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 601–610.
[35] P.F. Lazarsfeld, R.K. Merton, Friendship as a social process: a substantive and methodological analysis, Freedom Control Modern Soc. 18 (1954) 18–66.
[36] D. Lazer, A.S. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al., Life in the network: the coming age of computational social science, Science 323 (5915) (2009) 721.
[37] R. Leenders, Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection, Evolution of social networks 1 (1997) 165–184 Gordon and Breach.
[38] A. Lehrer, K. Lehrer, Antonymy, Linguist. Philos. 5 (4) (1982) 483–501.
[39] G.E. Lenski, Status crystallization: a non-vertical dimension of social status, Am. Sociol. Rev. 19 (4) (1954) 405–413.
[40] C. Leung, H. Chau, Weighted assortative and disassortative networks model, Physica A: Stat. Mech. Appl. 378 (2) (2007) 591–602.
[41] C. Levallois, Umigon: sentiment analysis for tweets based on terms lists and heuristics, in: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (* SEM), vol.2, 2013, pp. 414–417.
[42] J.R. Lincoln, J. Miller, Work and Friendship Ties in Organizations: A Comparative Analysis of Relation Networks, Administrative Science Quarterly 24 (2) (1979) 181–199 Sage Publications, Inc., Johnson Graduate School of Management, Cornell University.
[43] F.J. Massey Jr, The Kolmogorov–Smirnov test for goodness of fit, J. Am. Stat. Assoc. 46 (253) (1951) 68–78.
[44] M. McPherson, L. Smith-Lovin, J. Cook, Birds of a feather: homophily in social networks, Annu. Rev. Sociol. 27 (2001) (2001) 415–444.
[45] K. Meehan, T. Lunney, K. Curran, A. McCaughey, Context-aware intelligent recommendation system for tourism, in: Proceedings of the 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), IEEE, 2013, pp. 328–331.
[46] S. Milojević, Power law distributions in information science: Making the case for logarithmic binning, J. Am. Soc. Inf. Sci. Technol. 61 (12) (2010) 2417–2425.
[47] F. Mitzlaff, M. Atzmueller, A. Hotho, G. Stumme, The social distributional hypothesis: a pragmatic proxy for homophily in online social networks, Soc. Netw. Anal. Min. 4 (1) (2014) 1–14.
[48] M.L. Murphy, Semantic Relations and the Lexicon: Antonymy, Synonymy and Other Paradigms, Cambridge University Press, 2003.
[49] S.A. Myers, A. Sharma, P. Gupta, J. Lin, Information network or social network?: the structure of the Twitter follow graph, in: Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, pp. 493–498.
[50] M.E. Newman, Mixing patterns in networks, Phys. Rev. E 67 (2) (2003) 026126.
[51] M.E. Newman, J. Park, Why social networks are different from other types of networks, Physical Review E 68 (3) (2003) 036122.
[52] M.E.J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (2002) 208701, doi:10.1103/PhysRevLett.89.208701.
[53] H. Noel, B. Nyhan, The unfriending problem: the consequences of homophily in friendship retention for causal estimates of social influence, Soc. Netw. 33 (3) (2011) 211–218.
[54] R. Noldus, P. Van Mieghem, Assortativity in complex networks, Journal of Complex Networks 3 (4) (2015) 507–542 Oxford University Press.
[55] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (3) (1962) 1065–1076.
[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[57] M. Piraveenan, M. Prokopenko, A. Zomaya, Assortative mixing in directed biological networks, IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (1) (2012) 66–78.
[58] M.R. Piraveenan, Topological Analysis of Complex Networks Using Assortativity Ph.D. thesis, The University of Sydney, 2010.
[59] A. Portes, Social capital: its origins and applications in modern sociology, in: E.L. Lesser (Ed.), Knowledge and Social Capital, Butterworth-Heinemann, Boston, 2000, pp. 43–67.
[60] F.N. Ribeiro, M. Araújo, P. Gonçalves, M.A. Gonçalves, F. Benevenuto, Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods, EPJ Data Sci. 5 (1) (2016) 1–29.
[61] G. Rizzo, R. Troncy, Nerd: Evaluating Named Entity Recognition Tools in the Web of Data (2011).
[62] E.M. Rogers, D.K. Bhowmik, Homophily-heterophily: relational concepts for communication research, Public Opin. Q. 34 (4) (1970) 523–538.
[63] M. Rosenblatt, et al., Remarks on some nonparametric estimates of a density function, Ann. Math. Stat. 27 (3) (1956) 832–837.
[64] C. Roth, Generalized preferential attachment: towards realistic socio-semantic network models, in: Proceedings of the 4th International Semantic Web Conference and Workshop on Semantic Network Analysis, ISWC, 171, 2005, pp. 1613–1673.
[65] C. Roth, J.-P. Cointet, Social and semantic coevolution in knowledge networks, Soc. Netw. 32 (1) (2010) 16–29.
[66] H. Saif, Y. He, H. Alani, Semantic sentiment analysis of twitter, in: Proceedings of International Semantic Web Conference, Springer, 2012, pp. 508–524.
[67] S. Scepanovic, Implementation of ESA Algorithm for a Wikipedia SR Database, 2016, doi:10.5281/zenodo.49750.
[68] C.R. Shalizi, A.C. Thomas, Homophily and contagion are generically confounded in observational social network studies, Sociol. Methods Res. 40 (2) (2011) 211–239.
[69] S. Sun, H. Rui, Link formation on twitter: the role of achieved status and value homophily, in: Proceedings of the 50th Hawaii International Conference on System Sciences, 2017.
[70] J. Tang, H. Gao, X. Hu, H. Liu, Exploiting homophily effect for trust prediction, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ACM, 2013, pp. 53–62.
[71] L. Tang, Z. Ni, H. Xiong, H. Zhu, Locating targets through mention in twitter, World Wide Web 18 (4) (2015) 1019–1049.
[72] M. Thelwall, Heart and soul: sentiment strength detection in the social web with sentistrength, Proceedings of the CyberEmotions, 2013, pp. 1–14.
[73] C. Thurlow, L. Lengel, A. Tomic, Computer Mediated Communication, Sage, 2004.
[74] C. Wilson, B. Boe, A. Sala, K.P. Puttaswamy, B.Y. Zhao, User interactions in social networks and their implications, in: Proceedings of the 4th ACM European Conference on Computer Systems, ACM, 2009, pp. 205–218.
[75] L. Wu, Y. Ge, Q. Liu, E. Chen, R. Hong, J. Du, M. Wang, Modeling the evolution of users preferences and social links in social networking services, IEEE Trans. Knowl. Data Eng. 29 (6) (2017) 1240–1253.
[76] R. Xiang, J. Neville, M. Rogati, Modeling relationship strength in online social networks, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 981–990.
[77] M. Yavaş, G. Yücel, Impact of homophily on diffusion dynamics over social networks, Social Science Computer Review 32 (3) (2014) 354–372 Sage Publications Sage CA: Los Angeles, CA.
[78] X. Zeng, L. Wei, Social ties and user content generation: evidence from flickr, Inf. Syst. Res. 24 (1) (2013) 71–87.
[79] S. Zhang, M. van der Schaar, From acquaintances to friends: homophily and learning in networks, IEEE J. Sel. Areas Commun. 35 (3) (2017) 680–690.
[80] Y. Zhang, L. Leung, A review of social networking service (sns) research in communication journals from 2006 to 2011, New Media Soc. 17 (7) (2015) 1007–1024.

Sanja Šćepanović is pursuing her PhD in data science at Aalto Univ. in Finland, and, in parallel, the business and entrepreneurship degree at EIT Digital Academy. She analyzed large-scale datasets towards increasing the scientific knowledge in the fields of social networks, human mobility, Web cyber security and smart energy grid. She has been visiting research scientist at CERN and is also an alumna of International Space Univ. Her interest is data science in general, with particular focus on computational social science and the application of data science for the space domain.

Igor Mishkovski was born in Skopje, Macedonia, in 1981. He graduated and received the master degree in computer science and engineering at the University Ss. Cyril and Methodius, Skopje and the Ph.D. degree from Politecnico di Torino, Torino, Italy, in 2008 and 2012, respectively. After he received the Ph.D. degree in 2012 he was elected as an assistant professor at the Faculty of Computer Science and Engineering in Skopje. His research interests include complex networks and modeling dynamical processes, network science, computer networks, semantic web, operating systems.

Bruno Gonçalves is currently a Moore Sloan Data Science fellow at NYU's Center for Data Science while on leave from a tenured faculty position at Aix-Marseille Université. He has a strong expertise in using large scale datasets for the analysis of human behavior. Since 2008 he has been pursuing the use of Data Science and Machine Learning to study human behavior. By processing and analyzing large datasets from Twitter, Wikipedia, web access logs, and Yahoo! Meme he studied how we can observe both large scale and individual human behavior in an obtrusive and widespread manner. Main applications have been to the study of Computational Linguistics, Information Diffusion, Behavioral Change and Epidemic Spreading.

Trung Hieu Nguyen received the Ph.D. degree in computer science and engineering from Aalto University, Finland, in 2016. He is currently a postdoctoral researcher in the Faculty of Natural Sciences, University of Tampere, Finland. He has been a visiting researcher in the Next Generation Networks Group, University of Colorado at Boulder, between November 2015 and March 2016. His current research interests include distributed processing, cloud computing, energy-efficient cloud data centers, and big data analytics.

Pan Hui received his Ph.D degree from Computer Laboratory, University of Cambridge, and earned his MPhil and BEng both from the Department of Electrical and Electronic Engineering, University of Hong Kong. He is the Nokia Chair in Data Science and a full Professor in Computer Science at the University of Helsinki since September 2017. He is also a faculty member of the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology since 2013 and an adjunct Professor of social computing and networking at Aalto University Finland since 2012. He was a senior research scientist and then a Distinguished Scientist for Telekom Innovation Laboratories (T-labs) Germany from 2008 to 2015. He also worked for Intel Research Cambridge and Thomson Research Paris during his PhD period. He has published over 200 research papers with over 12,500 citations and has around 30 granted / filed European patents. He has founded and chaired several IEEE/ACM conferences/workshops, and has been serving on the organising and technical program committee of numerous international conferences including ACM SIGCOMM, IEEE Infocom, ICNP, SECON, MASS, Globecom, WCNC, ITC, IJCAI, ICWSM and WWW. He is an associate editor for IEEE Transactions on Mobile Computing and IEEE Transactions on Cloud Computing, a guest editor for IEEE Communication Magazine, and an ACM Distinguished Scientist.