

Semantic homophily in online communication: evidence from Twitter

SANJA ŠĆEPANOVIĆ, Aalto University
IGOR MISHKOVSKI, University Ss. Cyril and Methodius
BRUNO GONÇALVES, New York University
NGUYEN TRUNG HIEU, University of Tampere
PAN HUI, Hong Kong University of Science and Technology

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, show insights in its temporal evolution and finally, we present its intricate interplay with the communication network on Twitter. 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.

Keywords: Homophily, Semantics, Influence, Semantic Relatedness, Twitter, Wikipedia, Social Network Analysis, Computational Social Science

1. INTRODUCTION

Homophily [Lazarsfeld and Merton 1954; McPherson et al. 2001] (sometimes referred as selection [Leenders 1997; Crandall et al. 2008]) represents a tendency of individuals who are similar on some traits to connect to each other (become friends, follow each other, communicate etc.) in a social network. **Social influence** (peer pressure) is in a way an inverse tendency for people to become similar on some traits or to adopt certain behavior from their social contacts. Both, homophily and social influence result in a higher correlation (assortative mixing) on certain traits between connected than between random users in a network. This *assortative mixing* property (also in some studies referred as social correlation [Anagnostopoulos et al. 2008]) is repeatedly confirmed in social network analysis literature [Bollen et al. 2011; De Choudhury et al. 2010; Anagnostopoulos et al. 2008; Aral and Walker 2012; Tang et al. 2013]. 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 [Leenders 1997]. A third factor that could be the cause of social correlation is a common **external influence**. Moreover, a combination of these factors is often at play. For instance, an external factor might have non-homogeneous adoption in the network because friends could have a higher common latent propensity for it and adopt it to a larger extent than non-friends. Distinguishing between these factors as the main causes of assorta-

Author's addresses: S. Šćepanović, Aalto University, Department of Computer Science, Espoo, 02150, Finland; I. Mishkovski, University Ss. Cyril and Methodius, Faculty of Computer Science, Skopje, 1000, Macedonia; H. Nguyen Trung, University of Tampere, Tampere, 33100, Finland; P. Hui, Hong Kong University of Science and Technology, Department of Computer Science, Clear Water Bay, Kowloon, Hong Kong; B. Gonçalves, New York University, Center for Data Science, New York, 10003, U.S.;

tive mixing has been a challenge, and proven not even possible from observational data [Shalizi and Thomas 2011].

Extensive research is conducted in sociology on homophily in social networks as abstractions of diverse groups in society (see the seminal review by McPherson et al. [McPherson et al. 2001]). Classical paper [Lazarsfeld and Merton 1954] introduced two basic levels or dimensions 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.

In addition to individuals connecting to similar individuals, another suggested mechanism driving homophily is the process of *tie (link) dissolution* over time that happens more often among non-similar individuals. However, both mechanisms, of similarity and dissimilarity are not enough to explain a particular clustered (community) structure found in social networks. Sociologists have proposed that instead of only being driven by similarity, a tie is actually often formed around a specific **focus of homophily** [Feld 1981]. McPherson et al. [McPherson et al. 2001] offer a nice overview on possible different foci, and below we briefly discuss some of them. *Geographical proximity* is considered one of the most important foci of homophily, simply put, because we are more likely to have contacts with the people who are geographically closer to us. The ties induced by proximity in space are often weak; however, they leave more potential for stronger ties formation. It is worth noting that the advent of new technologies over time did not remove this pattern of geographical homophily and recent empirical research on online social networks finds that people online still tend to connect more often to geographically close people (in Twitter network [Kulshrestha et al. 2012; De Choudhury 2011]; in Microsoft IMS [Leskovec and Horvitz 2008]; in Facebook social graph [Ugander et al. 2011]; in mobile phone communication [Blondel et al. 2010]). The only study we found that reports no significant effects of geographical homophily tackles organization-individual relationship on Twitter [Sun and Rui 2017]. Another important focus that causes homophily are *family ties*. Family ties are an interesting focus of formation that causes people who are similar on some aspects and as well who are dissimilar on certain other aspects to connect. For this reason, when it comes to family ties, we find the largest geographic, age, sex and educational heterophily; but at the same time, the largest race, religious and ethnic homophily. *Organizational foci* turns to be the most important cause of ties that are not relatives nor family-bound. These foci include schoolmates, colleagues from work and voluntary organizations. A more implicit cause of homophily shows to be *network position*. 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 ties were random [Lincoln and Miller 1979]. Another, more internal, focus for homophily lies inside perceived similarity and shared knowledge, and it is termed *cognitive processes*. It is particularly notable among teenagers who tend to connect to those who are perceived to be more similar on some of the internal traits. Looking back at the described homophily traits and foci, it is not easy to make a clear distinction between the consequences of homophily and the causes or origins of it. Whether cognitive processes focus among teenagers causes them to become friends with similar ones; or whether the friend teenagers influence each other and hence become similar on a value homophily level?

1.1. Terminology

Communication network: In this study, the social network of interest is a *computer-mediated communication* [Thurlow et al. 2004] *network* from Twitter. It is formed of nodes representing Twitter users, and the directed links representing the tweets in which they mention (reply to) each other. The tweet content is also included. Hence, our network can be seen as a subtype of previously introduced *interaction networks* on Facebook [Wilson et al. 2009]. Throughout the rest of this study we simply use the term *communication network* referring to this definition. While in general communication refers to exchanging of information, we recognize the potential of Twitter mentions

to carry two different forms of communication. In the first form, the source is directly addressing the receiver, and 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)** in our network denotes the weight on the links i.e., the number of mentions between a pair of users.

Semantic homophily: Importantly, in many related studies the term *homophily* is used with the meaning of *assortative mixing* as we introduced it here (one possible reason being described indistinguishability of presented phenomena). We also use the term **semantic homophily** when talking about assortative mixing on semantic aspects of communication. In the light of introduced definitions, a more precise term to use 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. Hence, using an umbrella term semantic homophily to cover both tendencies is simpler. When at some point we talk about one of the tendencies in particular, we then point that out. In order to analyze semantic homophily, we tackle following *semantic aspects of communication*:

- **semantic relatedness (SR)** between the tweet contents of two users. SR is a more general metric compared to semantic similarity [Harispe et al. 2015] since in addition to similarity, it includes also any other relation between the terms, such as antonyms (opposite terms) [Lehrer and Lehrer 1982] and meronyms (a term is a part of or member of the other) [Murphy 2003]. 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 relation between tweets of a pair of users is quantified by a value ranging from 0 (not related at all) to 1 (maximally related);
- **sentiment** of user tweet content. The sentiment value ranges from -1 (negative) to 1 (positive);
- the most important **entities** (people, companies, organizations, cities, geographic features etc.), **concepts** (abstract ideas in the 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* [An IBM Company 2016]) and **taxonomy** (a hierarchy that helps to classify the content into its most likely topic category) of user tweets content.

Communication propensity (c_p) is defined as function of some property and represents 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 in our network with respect to SR threshold in the network (formula is given in Section 4.1).

Social capital: Among a variety of definitions from sociology [Portes 2000; Bourdieu 2011], one that translates well to our case introduces social capital as the actual and potential resources that are linked to the ego's social network and relationships. Hence, in similarity to the previous study on socio-semantic networks [Roth and Cointet 2010], we define social capital in our communication network as the total **number of contacts** (degree in the unweighted network) or the total **communication intensity** (degree in the weighted network). Moreover, we can divide the social capital, defined as such, in both cases to **popularity** (if we look at in-degree) and **communication activity** (if looking at out-degree). To sum up, thanks to our network being directed and weighted, we can introduce *four types of social capital* in it: (i) popularity in terms of number of communication contacts and (ii) popularity in terms of communication intensity and (iii) activity in terms of number of contacts and (iv) activity in terms of communication intensity.

Semantic capital denotes the amount of diversity of user tweet content with respect to the introduced semantic attributes, similarly as in [Roth and Cointet 2010].

Relative status of two users can be defined for both social and semantic capital and represents the difference of their respective status ranks. Finally, for a single user we define **status inconsistency** [Lenski 1954; Rogers and Bhowmik 1970] as a relative difference between his/her ranking among all users on social and semantic capital. Status inconsistent individuals tend to be highly ranked on some aspects and lowly ranked on others. This is suggested to be an attribute of individuals who are drivers of social change [Lenski 1954]. Status inconsistency can be defined on a communication

link, as well, as a measure of inconsistency of both participating users (we give a formal definition in Section 4.4).

1.2. Contributions

In this study, we offer a deeper understanding on the mechanisms of semantic homophily and how they are shaping the structure and properties of the underlying communication network.

While homophily has been identified in a diverse set of social networks, most of the studies investigated friendship, followers or citation type of ties. Interaction ties are more suitable for inferring meaningful social relationships [Wilson et al. 2009]. Our analysis is on the **communication ties** formed from Twitter mentions (replies), that are a subtype of interaction ties. The ties in our network are not only formed once (such as friendship and followership), but they require an active engagement over time. The nature of the mention network is fundamentally different from follower/friendship network in Twitter [Bliss et al. 2012]. For instance, the reciprocity of the followers network is found to be around 22% [Kwak et al. 2010] which is lower compared to the other social networks. The reciprocity of our mention network is 64%, considerably higher. When a user A follows a user B it simply states some type of potential interest in what B has to say. Depending on the different time zones and the number of other users that A is already following s/he might not even get to see any of B's tweets. In the case of our communication network we can clearly point to interactions and information diffusion between users (when the user A mentions the user B), instead of simply speculating about it when using the friendship/follower network. While retweet network allows for similar information diffusion analysis, its nature is also shown to be importantly different from mention network [Conover et al. 2011]. Finally, observance of communication interruption in time allows us to define a *tie dissolution (link decommission)*. As discussed in [Bliss et al. 2012], considering link decommission resolves issues of analyzing social network with stale links without current functional role.

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, through temporal evolution to its interplay with community structure of communication network.

Fig. 1 presents the general framework for our study and lists several main contributions. For a full list of our contributions, we refer the reader to Table VIII in Discussion 7. As depicted in Fig. 1, we operate on experimental datasets (from Twitter and Wikipedia), while at the same time building on existing sociological findings and theories. By testing for existence of status and value homophily, we confirm that these general theories from sociology hold in a communication social network. In addition, we identify the aspects of homophily that are specific for communication, compared to other types of social networks. A natural method to assess homophily in communication is through semantic aspects of it.

At first we **quantify** diverse aspects of semantic homophily in the network. We start by uncovering the subtle relationship between SR among users and the intensity of their communication. Next we introduce measures of social and semantic status of users and show that communication network exhibits assortativity on those metrics. This confirms sociological theories on status level homophily. We also show that such status correlation increases with strength of ties in communication. In addition, analysis of the interplay between two types of capital reveals large status heterogeneity among users. Accordingly, we find that status inconsistency of one or both communicating parties correlates with intensity of communication.

Next we focus on **temporal evolution** of semantic homophily. We detect temporal increase in average semantic relatedness among users and investigate new links formation as a possible cause. However, we also find a number of links that get decommissioned in time. After comparing relative statuses of users who stop communication, we present evidence that decommission is more

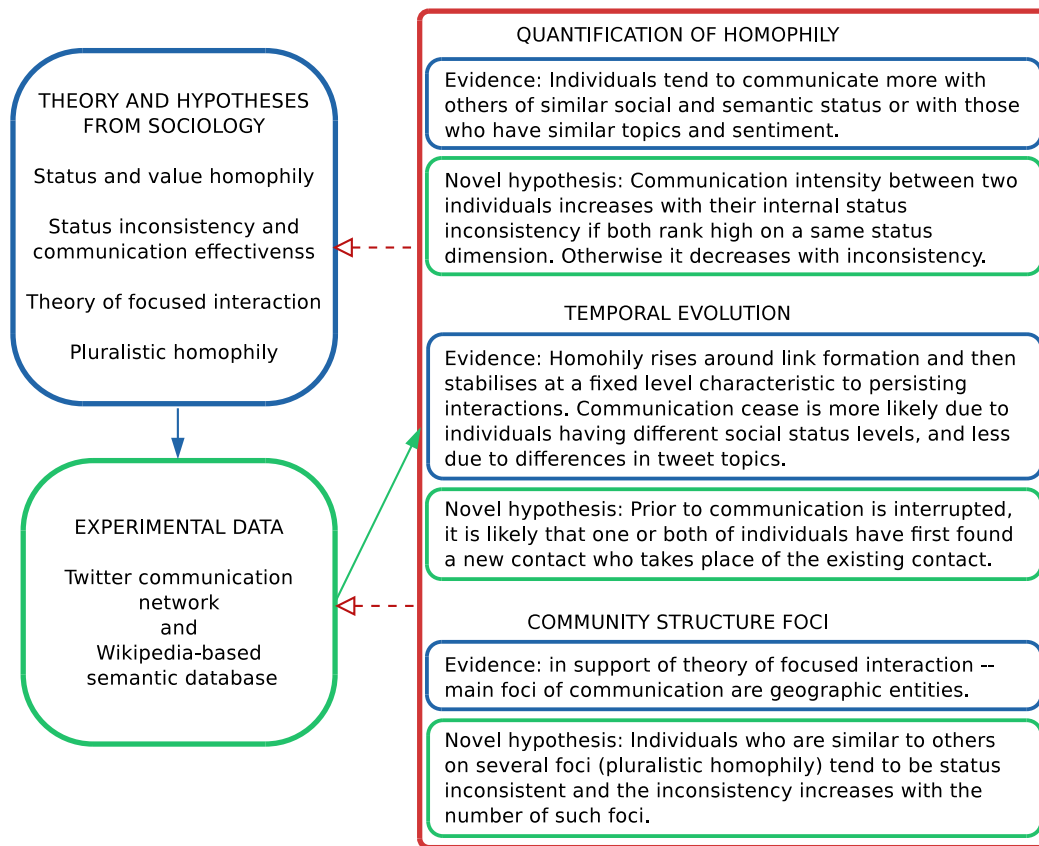


Fig. 1: **General framework and main contributions of our study.** In blue frames we denote the evidence found in Twitter experimental data for the existing theories from sociology. During data analysis, we also find evidence pointing to some novel hypotheses, presented in green frames. However, such evidence should be evaluated and confirmed in several other datasets before any general conclusions about semantic homophily in communication can be reached.

due to status than to value heterophily. Finally, the analysis on the **community structure** of the communication network (structural communities) reveals the semantic foci around which such communities are formed (functional communities). In this way, we find evidence for Feld’s theory of focused organization of social ties [Feld 1981] and also identify some of such foci around which communication ties are formed. In the end, we delve into the mechanisms of pluralistic homophily (assortative mixing as a result of several foci), and describe specificity of users who have such a position in communication network.

The rest of the paper is organized as follows. Section 2 presents related research literature. In Section 3 we describe two Web datasets (from Twitter and Wikipedia) used for analysis, as well as the framework of analysis consisting of a communication (Section 3.1) and semantic (Section 3.2) layer. Quantification of different forms of homophily in our network is presented in Section 4: social status homophily in 4.2 and semantic status and value homophily in 4.3. Insights on the relationship between these forms of capital, and relative status and status inconsistency are given in 4.4. The relationship between semantic relatedness and communication are reported in Subsection 4.1. Temporal aspects of semantic homophily, from link formation and dissolution to persisting interactions, are discussed in Section 5. Community structure and focused organization of social

ties are the topic in Section 6. Pluralistic homophily is also characterized in this section. The article concludes with a discussion and final remarks on future research directions in Section 7.

2. RELATED WORK

Knowledge networks representing scientific collaboration and blogger citations are studied in [Roth and Cointet 2010]. This study is similar to ours in that the joint dynamics and co-evolution of the social and socio-semantic structures is analyzed in these knowledge networks. Our work is different since we focus on another type of a network (communication). Hence, we respond in part to the call by Roth and Cointet [Roth and Cointet 2010] 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 are also able to investigate *link decommission (disconnection)*, thanks to the type of the network we analyze. Therefore, our work offers an additional understanding on the temporal interplay between semantic and social structures. Another important difference is that we offer considerably deeper semantic aspects analysis. Compared to a hand-picked set of categories used in [Roth and Cointet 2010], 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 analyzes homophily on the status (defined as the difference in the follower counts) and the value (tweet contents, common followees, location, age etc.) levels [Sun and Rui 2017]. There are several important differences to our work: the focus of their study is on reciprocal followers network (instead of mention network in our case), homophily is analyzed on the organization-individual relationship (whereas we focus on individual-individual relationship) and there is no focus on community analysis or temporal aspects of homophily as in our study.

However, there are several previous studies in online settings that have analyzed the *temporal interplay between homophily and social ties*. Crandal et al. [Crandall et al. 2008] find that the homophily between two Wikipedia admin users sharply rises some time before the tie formation and after that continues to slowly grow. This is interpreted so that, at first, homophily plays a role in the tie formation, but after that, the tie plays a role in the continuous increase of homophily. Another similar study on Flickr [Zeng and Wei 2013], 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 tie 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. Besides focusing on a different type of social ties – communication, our work extends these previous studies with the insights on interplay of homophily and *tie (link) decommission* that they have not investigated. In addition, we also uncover the relationship between introduced social and semantic forms of capital and homophily around the time of link formation and decommission.

Significant homophilous foci on Facebook [Barnett and Benefield 2015] are found to be geographic proximity, language, civilization, and migration. The analysis performed on 3 online datasets: Last.fm, Flickr and aNobii [Aiello et al. 2012], presents how homophily information can be used for link prediction. The authors present best accuracy in the case of aNobii (92%) when combining multiple features: in-degree, activity, number of distinct tags, assortativity of users in terms of topics etc. A conclusion is that the distinct language groups present in the aNobii dataset, which are quite homogenous and non-mixing, support the prediction accuracy. Halberstam et al. [Halberstam and Knight 2014] 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.

Below we mention several other studies that have tackled homophily in online settings, but with a different focus from us. A number of studies are conducted toward *distinguishing between influence and homophily* [Aral et al. 2009; La Fond and Neville 2010; Anagnostopoulos et al. 2008] report-

Table I: Twitter dataset filtering steps statistics

<i>Dataset</i>	<i>mentions</i>	<i>users</i>
original download	12 441 636	547 368
English language	2 527 990	284 100
users > 20 tweets	1 344 692	29 616
internal replies	744 821	26 717

ing different levels and proportions of the two traits in online social networks. For example, De Choudhury et al. [De Choudhury et al. 2010] 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, it is later shown that in empirical settings these tendencies are indistinguishable due to confounding effects [Shalizi and Thomas 2011]. 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 [Bakshy et al. 2012]. Hence this controlled experiment confirmed homophily or some unobserved common external influence taking place in the network.

3. DATASETS AND FRAMEWORK FOR ANALYSIS

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 [Bird et al. 2009] 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 knowledge database that we built (described in the following subsection), a threshold of minimum 20 tweets is selected. After this step, the dataset contains 29,616 users. Finally, again keeping only the internal replies within this group of users, we end up with 26,717 users in our final dataset for analysis (see Table I).

From the final filtered dataset we build our analysis target, the 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 II. Finally, at some points we will look at undirected and/or unweighted versions of the presented network. When we do so, it will be pointed out, otherwise, whenever we discuss communication network it refers to the weighted and directed network described here.

3.2. Semantic layers: Semantic enrichment of communication network

On top of the communication layer, we extract another, *semantic layer* from the Twitter data. Concretely, 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 relatedness**

Table II: Communication network statistics

<i>Network parameter</i>	<i>value</i>	<i>Network parameter</i>	<i>value</i>
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

database that we build from a whole English Wikipedia dump according to the Explicit Semantic Relatedness (ESA) algorithm [Gabrilovich and Markovitch 2009; Gabrilovich and Markovitch 2007]. The second procedure employs an existing, **natural language processing API, AlchemyAPI** [An IBM Company 2016] from IBM. Wikipedia SR database provides enrichment for both, edges (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. We describe both procedures and the enrichment they provide in more detail in the following.

3.2.1. Wikipedia Semantic Relatedness database. The semantic layer includes a network of users featuring semantic relatedness (SR) between their tweets collections as edge weights, we refer to it as the SR network. The SR network is based on SR knowledge database built using a Wikipedia XML dump from April 2015 (for details see Methods 8). In addition to SR scores, from the Wikipedia SR database, for each user we can also obtain their corresponding Wikipedia **concept vectors CVs**. CVs are formed of relevant Wikipedia concepts (articles) describing semantically user tweet contents.

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 communication network), resulting in a full SR network. Distribution of SR values of the full SR network is shown in Fig. 9 (right). 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.

3.2.2. AlchemyAPI. AlchemyAPI [An IBM Company 2016] 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 his/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 [Meehan et al. 2013] it was shown that the sentiment analysis obtained from AlchemyAPI achieved accuracy of 86% on a corpus of 5,370 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 [Rizzo and Troncy 2011] and [Saif et al. 2012], where AlchemyAPI outperformed Zemanta¹, OpenCalais², Extractiv³ and DBpedia Spotlight⁴ in extracting and categorizing named entities. However, besides the evaluations stated above, and the benchmark analysis done in [Ribeiro et al. 2016], we might consider using sentence-level methods, as VADER

¹<http://blog.zemanta.com/>

²<http://www.opencalais.com/>

³<http://extractiv.com/>

⁴<https://github.com/dbpedia-spotlight/dbpedia-spotlight>

[Hutto and Gilbert 2014], SentiStrength [Thelwall 2013] or Umigon [Levallois 2013] on our Twitter dataset as our future work.

4. QUANTIFYING SEMANTIC HOMOPHILY

4.1. Semantic relatedness and communication

We start by investigating interplay between SR for a pair of users and their CI by asking: *whether higher communication intensity is linked to a higher semantic relatedness?* Fig. 2 (left) displays the correlation when we apply logarithmic binning to account for long-tailed distribution of $CI(e)$. 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. Our result is comparable those in the study that evaluated similar relationship in retweet and follower Twitter graphs [Mitzlaff et al. 2014]. Next we turn to another way of assessing the interplay between the two communication aspects.

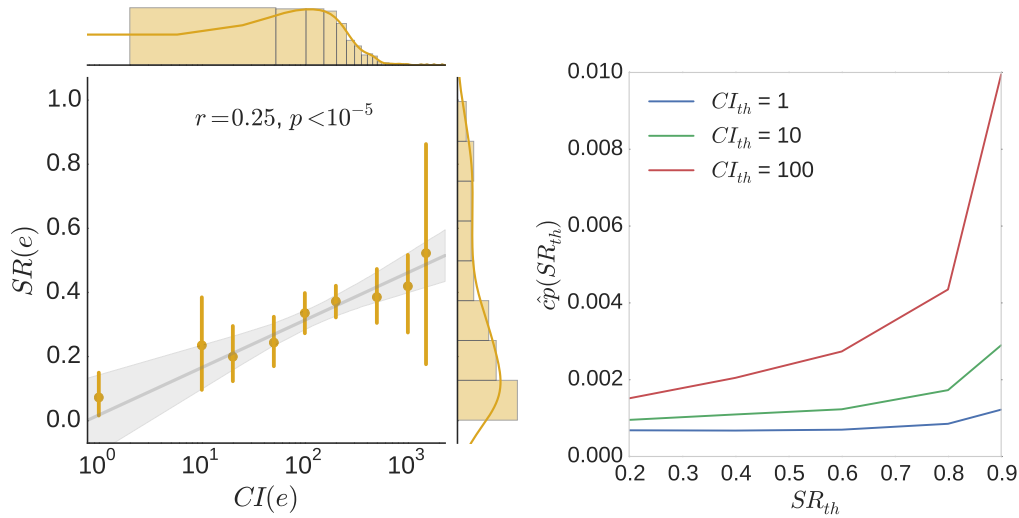


Fig. 2: **Interplay between SR and communication:** (a) correlation of link SR value ($SR(e)$) and its communication intensity ($CI(e)$); we apply logarithmic binning to account for long-tailed distribution of $CI(e)$; average value and standard deviation are shown for each bin; (b) communication propensity with respect to SR ($\hat{c}p(SR_{th})$) for different minimum communication intensity (CI_{th}) of links

We calculate **communication propensity** ($\hat{c}p$) with respect to SR threshold (SR_{th}) as the extent to which observed communication and its intensity diverge from what would be expected in a uniformly random setting. To illustrate, $SR_{0.2}$ network has $\sim 40M$ links, or $\sim 9\%$ out 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, we apply the dyadic propensity formula defined in [Roth 2005] to calculate $\hat{c}p$:

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

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. We also evaluate in the same

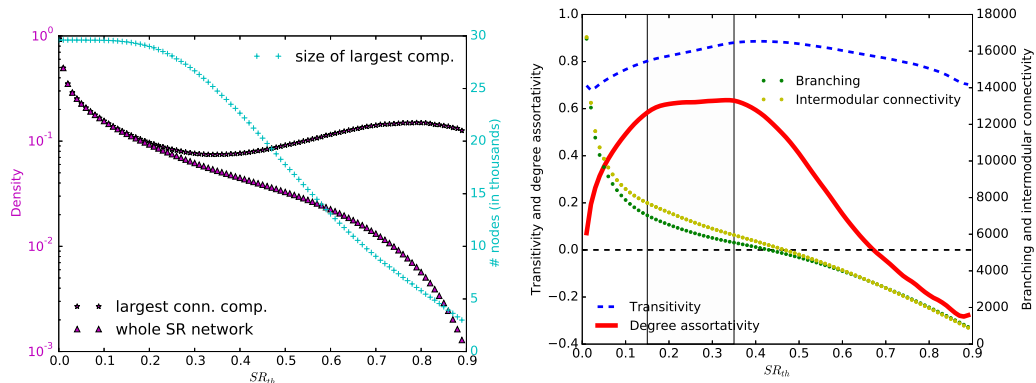


Fig. 3: **Properties of SR network in function of SR_{th}** : (left) size of the largest connected component, its density and overall network density; (right) branching factor, intermodular connectivity and transitivity as three ingredients for network degree assortativity [Estrada 2011]

way existence of links with a minimum communication intensity threshold (CI_{th}). Fig. 2 presents the results. Communication propensity increases with the increase in both SR and CI thresholds. The increase reveals presence of semantic homophily in the network with respect to SR . After both presented analyses, we conclude that the correlation between SR and CI is not simple and linear, but it is strongly captured by the subtle aspects of communication network.

This section we conclude with several results on the *properties of the full SR network*. It is important to point out that for this analysis we take the network built from the data for the whole 6 months period. Such results inform us about semantic relatedness metrics of a random group of people (not necessarily ever communicating). Fig. 3 (left) reveals that when thresholding the SR network near SR value 0.25, the largest connected component still has around 85% of the nodes and its density stabilizes, even it starts to grow, whereas the overall density in the network is significantly reduced.

In Fig. 3 (right) we plot the *degree assortativity* [Newman 2002] in SR network as a function of SR threshold. We detect an interesting changing pattern from positive to negative degree assortativity. In order to make sure that this pattern is specific to real-world SR metric, we randomize the SR values on SR network in several ways and find no pattern in such cases. Hence, we conclude that a structurally important change in human SR network takes place when we consider different SR threshold.

Fig. 3 (right) also shows the values for branching factor, intermodular connectivity and network transitivity (clustering coefficient), as it has been proven that they together define degree assortativity value [Estrada 2011]. In the interval $(0.15, 0.35)$ SR network obeys highest assortativity and transitivity. In this way we find lower and upper bounds for the threshold that can be used to remove the noise generated when building the SR knowledge database. For these values we also obtain the best community matching between SR network and communication network, as described in Section 6. From an application point of view, these findings might be important to consider while designing other semantic relatedness and similarity metrics, in particular when choosing a suitable threshold to distinguish significantly related and not related users.

4.2. Forms of social capital and degree assortativity

As we introduced earlier, a basic measure of assortative mixing in a network is the assortativity coefficient [Newman 2003] or simply assortativity. This coefficient is calculated as Pearson correlation between the value of a property on a node and the average value of that property on its neighbors. Hence the assortativity value ranges from 1 in a perfectly assortative network to -1 in a

perfectly disassortative network. Any discrete or scalar attribute of nodes can be used to calculate this coefficient.

We start by calculating assortativity based on node degree, an inherent node attribute of any network. Positive degree assortativity [Newman 2002] is suggested to be fundamental to social networks and to distinguish them from other types of networks [Newman and Park 2003].

Undirected network variants. We start by looking at an undirected variant of our communication network. Such an abstraction provides us with social capital in terms of number of contacts (unweighted) and total communication intensity (weighted network case). When we look at mutual edges, then we tackle *strong communication ties*, and when including all edges, then we also consider *weak communication ties* [Granovetter 1973]. The values of degree assortativity coefficient (r) in different variants of the communication network are presented in Table III. Using jackknife method as in [Newman 2003] we calculate and present also the standard deviation for each measurement to verify statistical significance of the results. Below we discuss and interpret the cases when our networks exhibits assortativity.

- Undirected unweighted network including all edges is **slightly disassortative** with $r = -0.015$.
- Undirected unweighted network with only mutual edges is on the other hand **highly assortative** with $r = 0.414$ (similar result reported in [Bliss et al. 2012]). This result shows that *the more strong contacts you have, the more strong contacts they themselves tend to have*.
- Undirected weighted network including all edges is **slightly disassortative** with $r = -0.014$.
- Undirected weighted network with only mutual edges is again **highly assortative** with $r = 0.474$. This result shows that *the stronger communication intensity you have, the stronger communication intensity your contacts tend to have*.

Directed network variants. In directed networks, four types of degree assortativity can be calculated, as introduced in [Piraveenan et al. 2012]. These four types of assortativity coefficients show if the degree of a source node is correlated with the degree of the target nodes, hence tackling relational analysis between source and receiver in communication [Rogers and Bhowmik 1970]. As shown in Table III the **in-in** is the only negative of the four coefficients in our network. This is in agreement with the findings for assortativity in directed followers Twitter network [Myers et al. 2014], except for **out-in** coefficient which is also found negative in the followers graph and it is slightly positive in our case. The authors (ibid.) 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.

- Looking at **in-in** coefficient, there was **no assortativity** with $r = -0.001$ in the unweighted network. This value increases to $r = -0.015$ in the weighted network case and becomes statistically significant. It is still low so we do not interpret it.
- Low positive **in-out** degree assortativity tells that: *the more popular you are the more active those who you contact tend to be (both in terms of number of contacts and in terms of communication intensity)*.
- Positive **out-in** degree assortativity is low (0.038) so we do not interpret it.
- The highest coefficient is for **out-out** degree assortativity, informing us that *the higher the number of users whom you contact, the higher the number of users they also tend to contact (or the more intensively you are communicating, the more intensively those who you contact also tend to be communicating)*.

Assortativity as a function of communication intensity. We can create an ensemble of weighted communication networks by thresholding the original network on different minimum edge weights. Then we calculate the above presented coefficients in each thresholded network. Since weight on the edges represents intensity of communication, the result is degree assortativity as a function of the communication intensity, as shown in Fig. 4.

Table III: **Degree** assortativity r coefficients in the communication network. Standard deviation s calculated using jackknife method [Newman 2003] is also presented

undirected networks		r	s	directed networks	r	s
unweighted	mutual edges	0.414	0.010	in-in	-0.001	0.002
	all edges	-0.015	0.001	in-out	0.110	0.013
weighted	mutual edges	0.474	0.017	out-in	0.038	0.003
				out-out	0.389	0.014
	all edges	-0.014	0.001	in-in	-0.015	0.002
				in-out	0.207	0.020
			out-in	0.014	0.004	
			out-out	0.338	0.026	

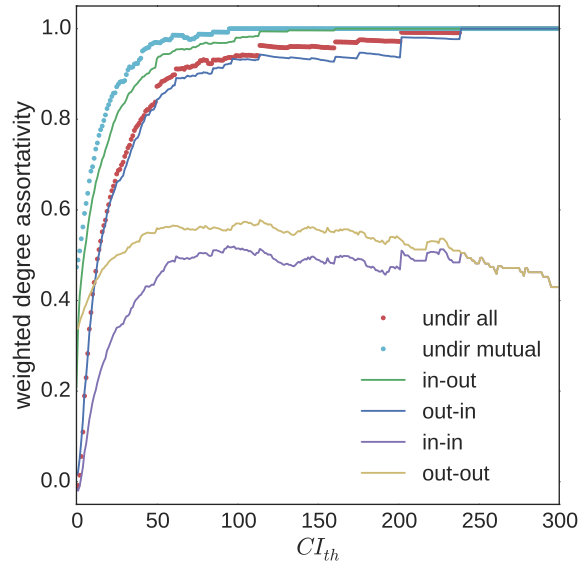


Fig. 4: **Degree assortativity** as a function of communication intensity CI in the ensemble of thresholded communication networks. *undir all* is degree assortativity in undirected network including all edges; *undir mutual* in undirected network with only reciprocal; *in(out)-in(out)* are the four types of coefficients in directed networks showing the correlation between *in(out)*-degrees of source and receiver nodes [Piraveenan et al. 2012]

First insight is that already with a small threshold, the two assortativity coefficients that are in the original network found slightly negative (in undirected network with all edges and in directed network **in-in** coefficient) become positive. With the threshold larger than 20 mentions, the networks are highly assortative on all the coefficients. This property exhibits one of the differences between often analyzed social networks based on unweighted, once formed links (such as friendship and followership) and the **weighted communication** network that we focus on. Bliss et al. [Bliss et al. 2012] demonstrated temporal stability of degree assortativity in mutual mention network, while herein we exhibit its variability with respect to the minimum communication intensity. Coming back to the above mentioned negative assortativity results in the Twitter followers network [Myers et al. 2014], we argue that at higher communication intensity (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.

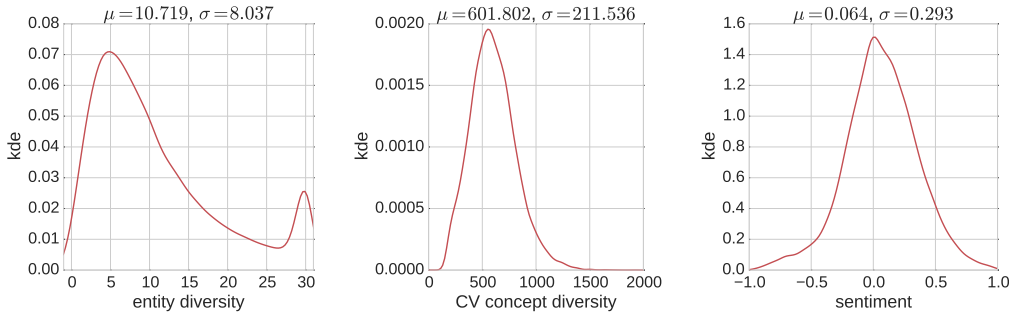


Fig. 5: **Semantic capital distributions**

Moreover, looking at the higher communication intensity thresholds, we notice two more interesting patterns. Two directed assortativity coefficients (**in-in** and **out-out**) start to slowly decrease, while the four other coefficients asymptotically reach the maximum value 1. In our concrete network case, the threshold of 239 mentions is when the four coefficients all become equal by reaching the value 1 and also the coefficients **in-in** and **out-out** become equal (at value 0.505). While not shown in Fig. 4, we calculated and those two coefficients continue to drop, while the others stay at the maximum value as we increase the threshold further.

To conclude, presented positive degree assortativity properties reveal presence of *social status homophily* (users with higher status tend to assortatively connect) on different forms of social capital in the communication network. We also find slight amounts of *social status heterophily* in relation to weak ties and popularity, but this heterophily quickly gives place to strong homophily when there is higher communication intensity in the network.

4.3. Forms of semantic capital and attribute assortativity

In this section, we investigate levels of assortative mixing on semantic aspects in the communication network. Besides degree, social networks are shown to exhibit assortativity on diverse nodes attributes [Bollen et al. 2011; Aiello et al. 2012; Eom and Jo 2014]. In line with such previous findings, we ask on which **semantic attributes** our Twitter communication network exhibits assortativity and to what extent. While social capital aspects presented in previous section reveal status homophily, some of the semantic capital aspects in this section exhibit value and some status homophily. Precisely, we look at assortativity on sentiment score and topics presence in the tweets, revealing *semantic value homophily*. We also look at semantic capital, or the diversity with regard to the number of relevant entities, concepts and taxonomy levels found in the tweets and this analysis reveals *semantic status homophily*. Prior to looking at assortativity, it is useful to familiarize ourselves with the distributions of semantic capital and sentiment values for the whole user base. In Fig. 5 we show kernel density estimates of their distributions displaying heterogeneity of entities and CVs diversity (see Section 3.2.1 and Methods for description of CVs), and sentiment 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 in their CVs, 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 sentiment scores). Hence, we conclude that there is large **semantic capital heterogeneity** among our users (see [Roth and Cointet 2010] for similar result in different types of networks).

Table IV: **Status and value homophily**: attributes assortativity r in the unweighted communication network. Standard deviation s calculated using jackknife method is also presented

<i>level</i>	<i>status homophily</i>				<i>value homophily</i>			
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

The results presented in Table IV suggest the presence of both, **value** (topics of tweeting, sentiment) and **status** (semantic capital) homophily in the unweighted versions of the communication network. We focus on the unweighted versions, since we first of all ask, whether there is a tendency among the users to have contact with other users who are similar to them on some semantic attributes (without looking at intensity of communication). This means that the answers to this question in the networks including only mutual edges will inform us about such correlation among strong contacts, while looking at networks with all edges included will inform us also about weak contacts. Once again, as with the degree assortativity, we find that mutual (reciprocal) communication network is importantly different compared to the network including also one-sided communication edges. Notably, it exhibits higher levels of assortativity on all the analyzed attributes.

As the observed correlation levels could be induced by existing degree assortativity, we also test the presence of assortativity after node attribute randomization. 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 establishing the presence of status and value homophily in the communication network on different forms of social and semantic capital, we ask next about the relationship between these forms of capital. Whether the users who are richer in terms of social capital (and hence more network central) are also richer in terms of semantic capital (their tweets are semantically richer, or exhibit more diversity on semantic aspects)? With this analysis, we respond to the call by authors in [Roth and Cointet 2010] to look for similar types of patterns as they have investigated in the bloggers and scientists networks. Indeed, we also find a wide range of possible combinations of joint values of social and semantic capital, as they have reported. In the end we conclude that the observed patterns in the Twitter communication network resemble more of the bloggers than the scientists network presented in [Roth and Cointet 2010].

Precisely, testing for different forms of social capital against different forms of semantic capital reveals no significant or low to medium correlations between the two. For the purpose of visualization, in Fig. 6, we show joint distributions for entity, concepts diversity and sentiment score on one side and communication intensity (of popularity and of activity) on the other.

When it comes to **popularity** (weighted indegree), we observe a wide spectrum of semantic diversity in terms of entities for both, the users with low and high popularity. Most popular users tend to be slightly more likely to have high semantic diversity. On the other hand, most popular users are likely to have quite neutral sentiment in their tweets. However, users which are more positive or negative in their sentiment are likely to have modest to low popularity.

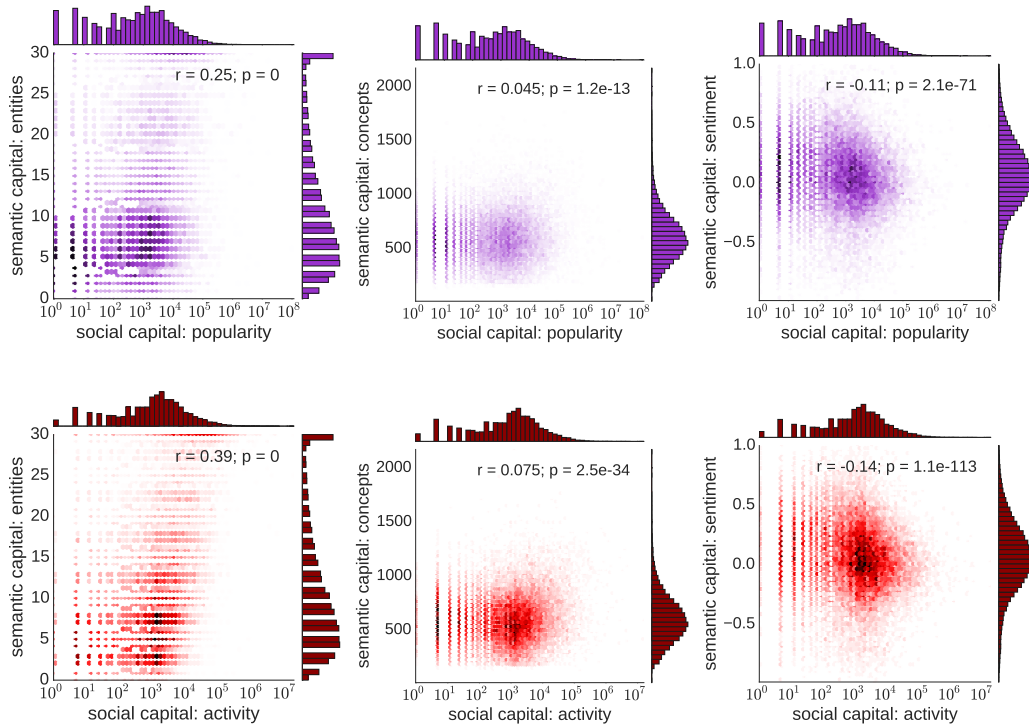


Fig. 6: **Joint distributions of social and semantic capital:** the darkness of the hexagon corresponds to the frequency of users with the combination of social and semantic capital values

When it comes to intensity of **communication activity** (weighted outdegree), we observe similar patterns that are a bit more pronounced for the socially richest users. Basically, most actively communicating users are likely to have higher semantic diversity in terms of entities (however, we still find a number of users with diverse tweet contents that are not actively communicating). Semantic (entity) diversity has the highest correlations with communication activity (weighted outdegree; $r = 0.397$) presented in Fig. 6 and with weighted mutual degree ($r = 0.396$). These values are similar to the value found in the bloggers network and lower compared to the scientists network in [Roth and Cointet 2010].

Sentiment has *negative correlations* with both popularity and activity, also presented in Fig. 6. This means that with popularity and being active users tend to have a slightly more negative tweets sentiment. Finally, when it comes to diversity in terms of number of concepts present in their CVs, 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). Hence, we conclude that different forms of semantic capital have different patterns of interplay with social capitals.

Thanks to our network being directed and weighted, we are able to observe one additional pattern: while being particularly low for popularity (indegree), all the correlations increase for user activity (outdegree) and with communication intensity (weighted degrees). For instance, the correlation between entity diversity and (unweighted) indegree is only 0.051. In this way, we exhibit that *communication activity, intensity and stronger contacts* are more conducive of *higher semantic capital*, compared to popularity and weaker contacts.

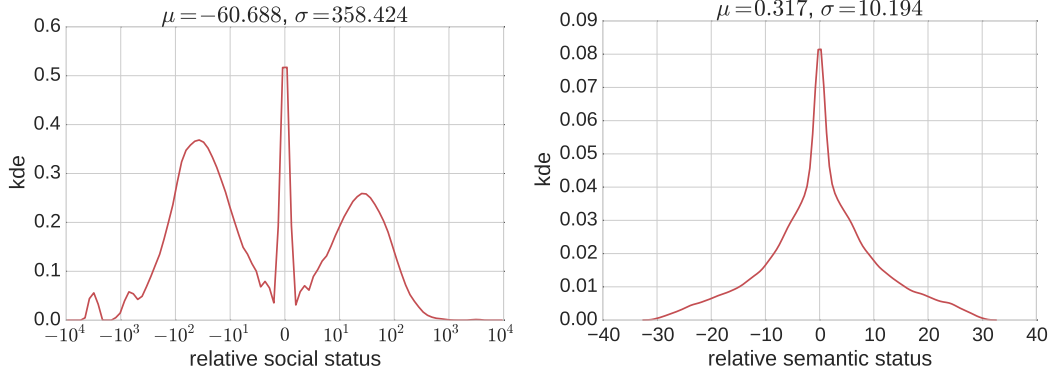


Fig. 7: **Status differences in communication:** Kernel density estimates for distributions of (left) popularity difference and (right) semantic capital difference.

Relative status of source and receiver. An additional way to investigate the interplay between social and semantic capitals is in terms of **relative status** of source and receiver in communication. By relative status we mean the difference in status on a particular form of capital. Such definition is similar to the achieved status presented in [Sun and Rui 2017]. In Fig. 7 we show distributions of relative social status (popularity difference) and relative semantic status (entity diversity difference) between source and receiver. In particular, the distribution for relative social status exhibits a dominant peak at zero (users with similar status are most likely to communicate), but plotting it 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 social status belonging to these ranges to be talking to each other. The left peak is higher, and this together with the negative mean value for relative social status informs us 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 semantic capital, most of communication happens between those who have close to equal semantic capital.

For the joint distribution of social and semantic relative statuses we find (analyzed, not shown in a graph) 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*.

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 [Rogers and Bhowmik 1970]. *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* [Lanski 1954]. 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 **status inconsistency** (st_{inc}) as in [Lanski 1954]:

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

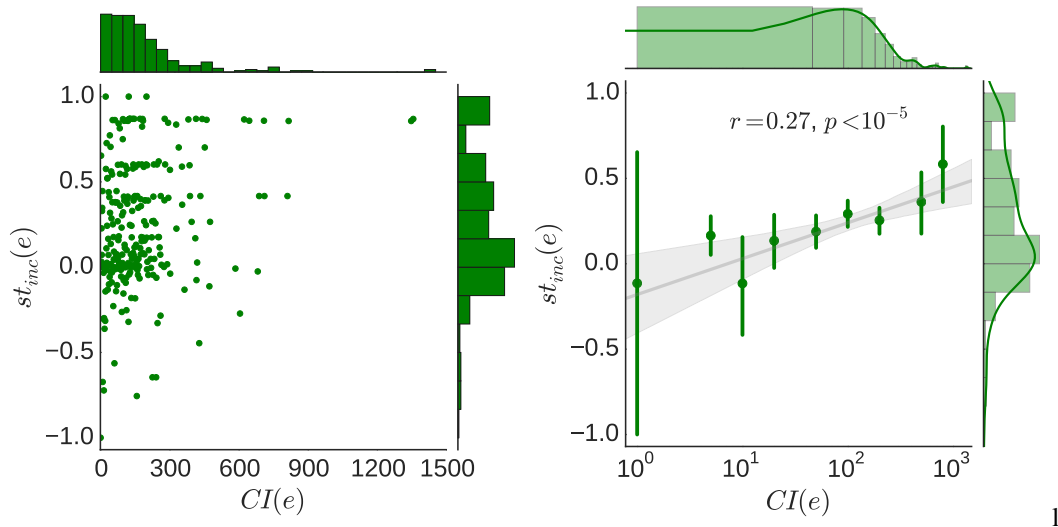


Fig. 8: **Relationship between communication intensity and link inconsistency:** (left) scatter plot; (right) linear regression visualization – we apply logarithmic binning to account for long-tailed distribution of $CI(e)$; average value and standard deviation are shown for each bin.

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 assess the amount of user status inconsistency (how close is $abs(st_{inc})$ to 1), and second, it also encodes whether he/she has higher social (st_{inc} is positive) or semantic (st_{inc} is 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. Now, for all the directed links $(e_{i,j})$ in our communication network we define **link inconsistency** using above introduced status inconsistency of the source ($st_{inc}(u_i)$) and the receiver ($st_{inc}(u_j)$) as their product:

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

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

We indeed find significant correlation between introduced link inconsistency and communication intensity ($r = 0.27$). Results presented in Fig. 8 indicate following finding: the communication between two users tends to increase with status inconsistency of one or both of the users, if they are both richer on the same form of capital. If the users are status inconsistent but being rich on different forms of capital, then their communication intensity tends to decrease. As with other findings regarding social capitals, the described patterns are relevant for extreme cases (high and low edge weights), and there is a wide spectrum of edge inconsistency values taken by the medium-weight edges (Fig. 8, left).

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

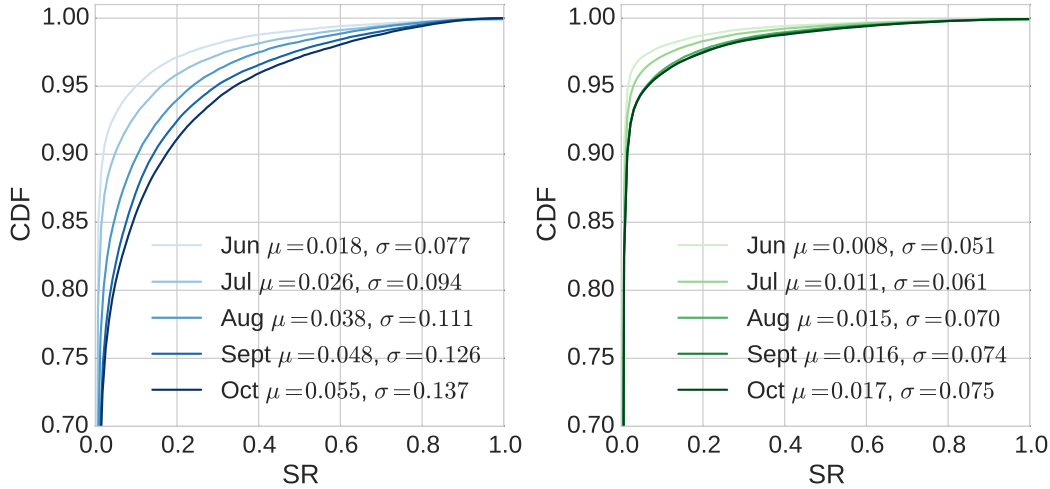


Fig. 9: **Cumulative SR distributions** for 5 full months in our dataset: (left) in communication network and (right) in the rest of SR network. For better visualization of the differences in distributions the y-axis is thresholded above 0.7. The distributions are together sharply rising up to around that point.

at different snapshots of the network for each month. First we analyze *temporal change of SR values*. In Fig. 9, cumulative distribution functions (CDF) of SR values for each full month in our dataset are shown for communication network and for the rest of the links in SR network. Precisely, we consider all the links with mutual communication (strong ties) in communication network, while for the second distribution, we take the difference between links in SR network and all communication contacts (both strong and weak). In this way we aim to distinguish between SR of user pairs affected by communication (and hence social influence) and those that are less likely to be affected (no communication of any type occurred between them in our dataset). Gradual increase in SR values takes place in both cases over time (CDF increases at higher SR values). In addition to the visualization, by applying Kolmogorov-Smirnov (K-S) test [Massey Jr 1951] we confirm the distribution change. In particular, we compare the distributions for June and for October. For communication network, K-S results in $p < e^{-24}$ and, respectively, for SR network, in $p < e^{-197}$, hence in both cases strongly rejecting the hypothesis that the distributions are the same.

5.1. External influences evidence

The increase in average SR in SR network (Fig. 9) among not connected pairs of users is peculiar. It indicates a possible external influence taking place during the period causing all users to talk more on a similar (external) topic. However, since the Twitter social network we investigate is not the only possible way 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. In any case, we turn to our semantic layers to look for an evidence of common external influences in the dataset.

Using **AlchemyAPI** output, we identify overall most popular categories for topics of communication in our dataset. They are displayed in Fig. 10. *Arts and entertainment*, including *movies*, *tv shows*, *music* and *humor* is the dominant category. Second set of most popular categories includes *sex* (under *society*), *sports* and *technology and computing*.

Insights on common topics of communication using **Wikipedia semantic relatedness database** are consistent with those from AlchemyAPI. In Table V we present some of top 100 concepts

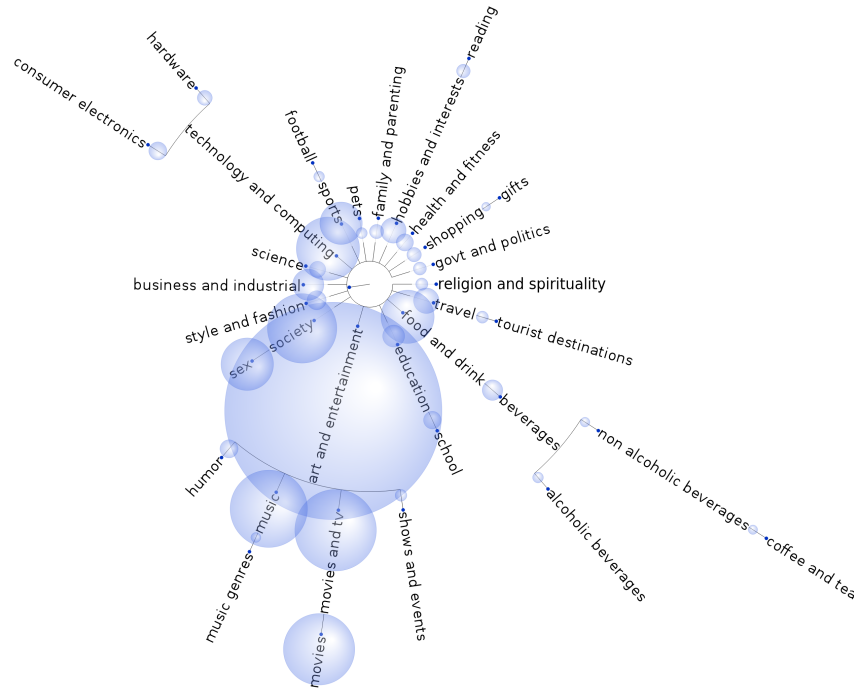


Fig. 10: **Semantic taxonomy of whole communication network visualized in a bubble-tree-map**: highest level categories are in the center. Subcategories are represented as descendants in the tree. Size of bubbles corresponds to the frequency of topics under that category in our dataset.

(Wikipedia articles) found to describe the semantics in the dataset overall. For easier comparison, we display these concepts per (sub)categories identified using AlchemyAPI. The two seasons of TV series *This Is England* that have been aired at the time corresponding to our dataset are ranked 2nd and 3rd. Next, we also find several musicians and bands. The concepts *LOL* and *Smiley Face* are in part a result of how ESA algorithm [Gabrilovich and Markovitch 2009; Gabrilovich and Markovitch 2007] that we used to build Wikipedia SR database works. They are also in agreement with humor being prevalent subcategory among users in our dataset. In addition to the series *This is England* being aired at the time of our dataset, the death of Osama bin Laden also happened during that period, and we see an article about him describing the general conversation. ESA’s output of > 300K Wikipedia concepts describing topics in our dataset results in a fine SR metrics, as exhibited in detecting fine gradual temporal increase. At the same time, from Table V we see that already the top 100 concepts provide insights into the concrete topics of the conversation in the dataset.

These insights, offer evidence for some external influence taking place in our dataset that could lead to global increase in SR among not connected users. Since mentioned TV series, music and events are prevalent topics in the dataset, it could mean that our users are independently watching/following and commenting on them. This in turn could lead to average increase in their SR, even if they never communicated. However, once again, we can not assert whether the increase is indeed (only) due to external influence or due to some social contacts and/or peer influence not detectable using our Twitter dataset.

Table V: Most popular Wikipedia concepts in the dataset, per taxonomy categories: movies and TV shows, music, sports and humor

<i>Wikipedia articles in category</i> <i>Movies and TV shows</i>	<i>Concept rank</i>	<i>Wikipedia articles in category</i> <i>Music</i>	<i>Concept rank</i>
This Is England '86 (TV series)	2	Robert Smith (musician)	5
This Is England '88 (TV series)	3	10cc (English rock band)	9
Love of Life (American soap opera)	15	The Cure	10
The Dad Who Knew Too Little (Simpsons episode)	38	Producers (band)	16

<i>Wikipedia articles in category</i> <i>Sports</i>	<i>Concept rank</i>	<i>Wikipedia articles in category</i> <i>Humor</i>	<i>Concept rank</i>
List of electronic sports titles	22	LOL	1
Larry Johnson (American football)	67	Smiley Face	4
Alabama Crimson Tide football	68	Lolcat	20
Racism in association football	82	Pres. Obama on Death of Osama bin Laden (spoof)	36

5.2. Semantic homophily, social influence and tie dissolution

The increase in communication network can be due to homophily in its strict definition, i.e., new user pairs starting communication. Once connected they are later likely to have higher SR, as we presented in Section 4. This can happen due to already connected pairs becoming more related, i.e., social influence. Sociology also suggests to look for link dissolution among dissimilar individuals [Felmlee et al. 1990; Block and Grund 2014] as one of the reasons of average network SR increase.

We start by investigating formation and dissolution of links through time and their SR change. 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 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.

Temporal change of average SR on links prior to and after the *activation* is shown in Fig. 11. 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 [Crandall et al. 2008] and for Flickr users [Zeng and Wei 2013]. The drop in average SR in the period after the link activation is also reported in earlier studies [Zeng and Wei 2013]. 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 formal edge formation (adding someone as a friend or following) and do not require an active user engagement afterwards.

Indeed, we find in total 8,166 links that are *activated and then also decommissioned* during our dataset period. The SR change for such links, that are activated and then decommissioned, as well as for those that persist in our dataset after the formation is show in Fig. 12. The average SR values for formed and persisting links stay high after they are formed. It is those links that will get

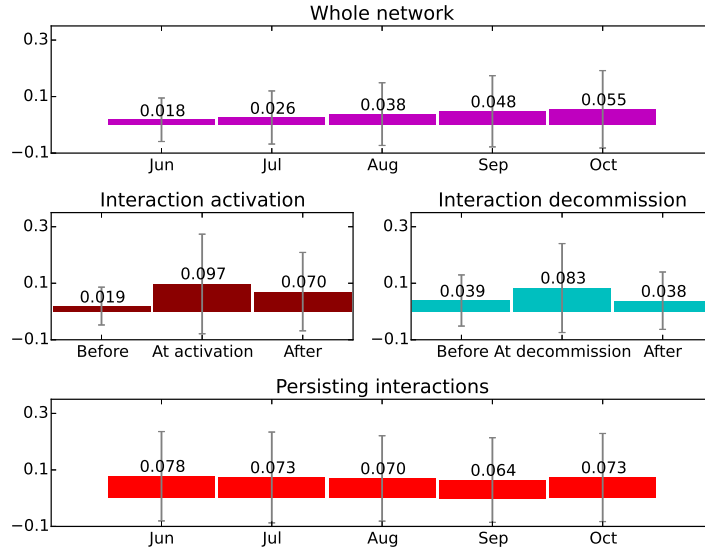


Fig. 11: **Temporal SR change.** Average SR on: (top) all communication links, (mid, left) during communication activation, (mid, right) decommission, and (bottom) on persisting links; error bars show standard deviation values

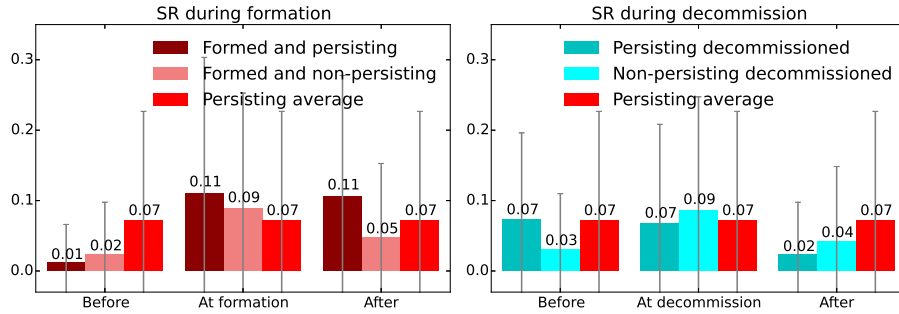


Fig. 12: **SR change during formation and decommission;** (left) during link formation and (right) during link decommission. We show the differences between the links that persist after formation or not, and similarly between those that were persisting in our dataset period before decommission and those that were non-persisting. Error bars show standard deviation values

decommissioned soon that contribute to lowering the average SR after formation that we see in Fig. 11. This result displays that homophily needs to be considered together with active engagement and its temporal dynamics.

If observing only the *persisting links* that were already active and persisted during the whole period in our dataset, we obtain results for their average SR change in the bottom plot in Fig. 11. Such persisting interactions have 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

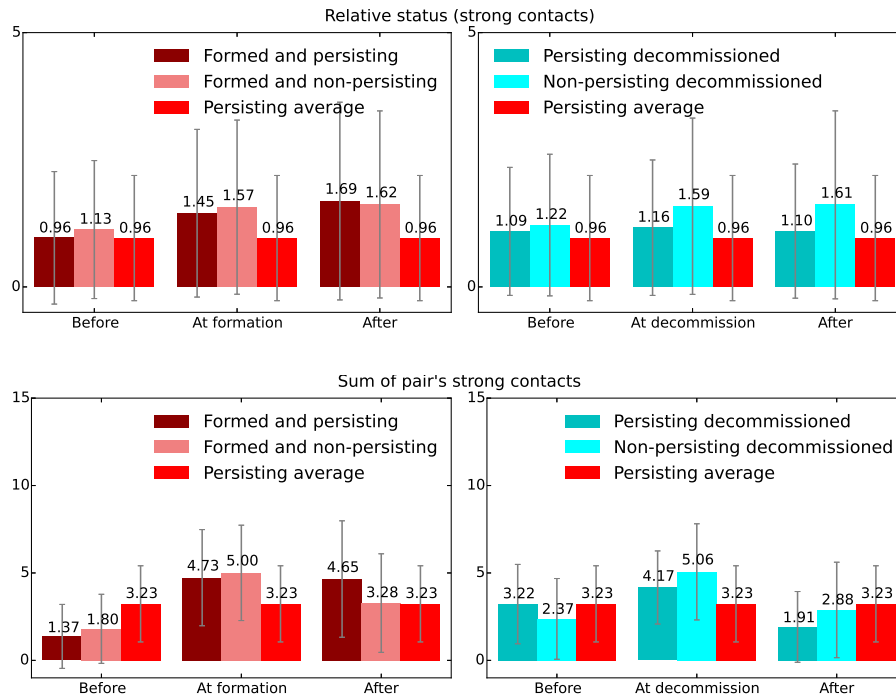


Fig. 13: **Temporal status differences during formation and decommission:** (top) average relative social status (number of strong contacts), (bottom) average sum of strong contacts. Error bars show standard deviation values

which persist and have high SR, 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, as indicated by average SR during dissolution of previously persisting links in Fig. 12 (we discuss this result in more detail below).

Fig. 11 also displays temporal change of SR on links that get *decommissioned*. Again, in Fig. 12, we separate persisting links (during our dataset) that get decommissioned from those that have formed during our dataset time frame (non-persisting) and get decommissioned. Indeed, we can 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 more stable relationship between 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. Sociology suggests as one possible cause for link decommission that maintaining ties with dissimilar others might be costly [Felmlee et al. 1990]. However, we notice that the SR values before decommission on previously persisting links are not lower but around the same as on the links that stay persisting. Hence, *in terms of SR there is no observable dissimilarity between users with persisting communication before they will cease communication*. We investigate other possible reasons for their link dissolution below.

Operating on the same sets of communication 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 mutual communication, it is natural to assess social capital in terms of

numbers of strong/weak contacts. In Fig. 13, we show relative status and total number of strong contacts of 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. Looking at relative status (top row plots in Fig. 13), we first notice the difference on persisting links compared to other types of links (and also to the whole network, a result which is not displayed). Persisting links have lower relative status, i.e., users who are actively communicating tend to have similar social status rank. While homophily on the status level is not new, herein we exhibit its underlying mechanisms in communication network. Namely, both types of links, those that are newly formed and those that will get decommissioned in time, 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). The results are similar for relative status in terms of weak contacts so we do not present them. To reiterate, our analysis so far gives two insights about links before they get decommissioned: i) lack of semantic differences on previously persisting links (their SR is not lower at the time when link dissolution happens compared to those who consistently persist communication) and ii) higher status differences (also compared to persisting links). Hence, there is indication in Twitter network that *persisting communication links dissolve in the presence of status level heterophily rather than value level heterophily*.

Findings from sociology also suggest that relationships last shorter time and are more likely to decay for pairs of individuals with lower overall *social status* [Burt 2000]. To assess this hypothesis in our communication network, we observe social status in terms of total number of contacts for user pairs who cease communication. Results in Fig. 13 (plots in bottom row) do not support such hypothesis for strong contact: pairs who cease communication have around the same sum of strong contacts on average as the pairs who persist communication. Moreover, in the case of weak contacts, there is an opposite evidence: pairs prior to communication cease tend to have more weak contact compared to average of persisting links (other results for weak contacts are similar to strong so we do not show them). In addition, the increase in the sum of pair's contacts at the month of decommission suggests that those new contacts might affect their existing link. After the decommission the sum of contacts drops, but still stays higher than would be expected after the decommission (existing link is counted as one strong contact for both users, so after the decommission, their sum of contacts would be expected to drop by 2). Such evidence suggests that in some percent of the cases *one or both of the users have established new communication links at the time of abandoning the current one between them*. This result is supported by a level of stability on the number of persisting communication links per user. Namely, most of the 5,229 users who participate in constantly persisting links in our dataset have between one to two persisting contacts ($\mu = 1.2$ and $\sigma = 0.49$).

In summary, presented types of interactions show the importance of considering both homophily and influence as dynamic interdependent tendencies [Yavaş and Yücel 2014] in temporal networks, instead of looking at static snapshots. Our analysis on interaction decommission reveals similar results as in [Noel and Nyhan 2011] where it is showed how not accounting for homophily effect on tie dissolution ('unfriending') 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 as a future work, we aim to look at the period in which edge formations and deletions might be happening, and whether there are some natural cycles in the human communication networks.

6. COMMUNITY STRUCTURE AND SEMANTIC FOCI

We start by investigating what are the semantics traits that shape community structures in communication network of Twitter users.

Table VI: Largest communities in the communication network and their semantic foci

<i>Num of users</i>	2222	686	636	435	381	343	343
<i>Main geo-entities</i>	Nigeria	Indonesia	South Africa	Philippines, Malaysia	Jamaica	U.K.	NY, LA, Miami
<i>% positive users</i>	0.38	0.87	0.67	0.72	0.5	0.59	0.71

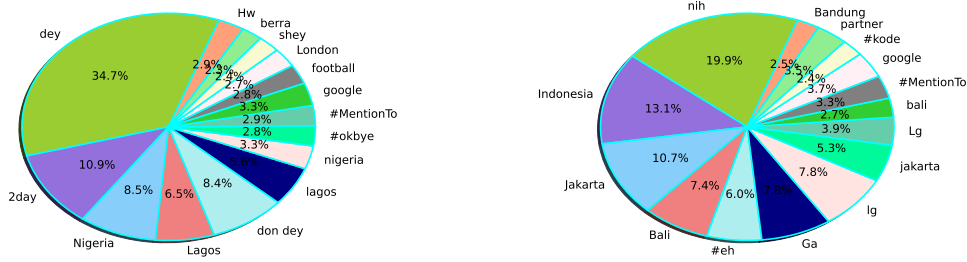


Fig. 14: Most relevant entities found in the tweets of the two largest communities: (left) Nigerian and (right) Indonesian

When dealing with representations of real-world networks one can distinguish between structural and functional communities [Yang et al. 2014; Yang and Leskovec 2015]. The connectivity pattern among members in the network defines **structural communities**, whereas a common function or a role of user groups defines **functional communities**. Simply speaking, structural communities can be defined as groups of users that are more tightly connected within the group compared to the rest of the network. This definition can entail **modular** or communities with distinct users, but also, more representative of the real-world, we can think of **overlapping community structure**, where certain nodes belong to more communities.

If we recall Feld’s theory about *foci of homophily* [Feld 1981] that drive clustered (community) structure of social networks, then foci can be seen as one such common function or role around which communities are formed. In our case, we allow different semantic traits of user communication to define semantic foci. Our initial question can be now rephrased as *whether structural communities (both modular and overlapping) can be explained in terms of their functional roles by semantic foci*.

6.1. Modular communication communities

A state of the art algorithm when it comes to detecting **modular community structure** is based on modularity metrics [Newman 2006]. We run its fast implementation [Blondel et al. 2008] on our communication network and detect 2632 communities. Statistics about the largest detected modular communities is shown in Table VI. By applying semantic analysis on groups of users belonging to detected communities, we find most relevant semantic traits of the communication in each community. Precisely, we find relevant concepts, entities, categories, taxonomy tree and average sentiment for each community. Then we also apply TF-IDF analysis on the semantic traits with respect to those for the whole communication network to assess whether found semantic traits are specific to a community. After careful analysis, we conclude that only the entities of conversation can be used to explain the modular communities. As an example, in Fig. 14, we present top entities

found in tweets of the two largest communities. Thanks to those entities, we are able to conclude that they represent respectively a community of users speaking about Nigeria and about Indonesia. Importantly, in addition to a few dialect specific words (such as in this case *dey* in Nigerian and *nih* in Indonesian community), among most relevant entities we find geographical entities (in addition to *Nigeria*, we find entity *Lagos* in Nigerian and in addition to *Indonesia*, we detect *Jakarta* and *Bali* for Indonesian community). With such analysis and additional manual inspection of the tweets, we conclude that the largest modular communities are formed around **geographic entities** as foci of communication (see Table VI for the other top size communities). To reiterate, we conclude that *geographic entities are homophilous foci that best explain modular communities in our communication network*. Similar result are found in different types of communication networks; good predictors of cohesive communication groups in [Leskovec and Horvitz 2008; De Choudhury 2011] are geographic foci and several studies [Blondel et al. 2010; Aiello et al. 2012] report language foci. As a remark, the communities in our Twitter network may be formed due to the ethnicity of users or their geolocation, while in any case, their tweet contents contain relevant geo-location entities.

Another important finding regarding modular communities is that there is a wide diversity in their average **sentiment**. In Table VI, we show the percent of 'positive' users in the whole community. We can see it ranges from 0.38, for a quite 'negative' Nigerian, to 0.87 for the most 'positive' Indonesian community. The large difference in the sentiment between these two particular communities can be also inferred from their relevant concepts: prevalent swear word-concepts in Nigeria (having negative sentiment), and, on the other hand, *gratitude* and *luck* being dominant in positive Indonesia. Displaying particularity of the Indonesian community, an earlier study found that Indonesian users have higher than average tweets per user ratio, which is related to higher reciprocity, and in turn a higher-reciprocity communities display a happier language [Poblete et al. 2011].

If modular structural communities were not formed around foci as suggested by Feld's theory, but if instead they were simply a result of semantically related users connecting more often, then we would expect to see similar communities when running community detection on the SR network. We test such *hypothesis* by detecting communities on a several SR_x networks. In order to evaluate how well the sets of communities from communication (*P*) and semantic layer (*L*) match, we apply the procedure used in [Yang et al. 2014; Yang and Leskovec 2012; Yang and Leskovec 2015] to find the matching score:

$$S = \max_{P_j \in P, L_i \in L} F_1(L_i, P_j),$$

where $F_1()$ uses F_1 as a score for similarity between the two sets. Resulting $S \in [0, 1]$, where 1 indicates perfect matching.

Best matching score we find when running InfoMap algorithm [Rosvall and Bergstrom 2008] on the SR_{0.2} network. The threshold $x = 0.2$ matches with and is explained by the analytical analysis of SR network that we discussed earlier (see Section 4.1). InfoMap is not modularity-based community detection and the rationale why it performs better on the *semantic layer* is because SR_x networks are so dense. Modularity metric, which is optimized by modularity-based algorithms, evaluates existence of dense connections among nodes within communities but **sparse** connections with nodes in different communities. Hence it can not work well on dense networks, such as SR_{0.2}. Best matching scores for biggest communities (with more than 50 users) are presented in Table VII. We also visualize modular communication communities and their respective SR community counterparts in Fig. 15. The matching scores reveal that SR communities can only to a moderate extent explain the communication community structure. Such conclusion, in turn, supports Feld's theory about foci of homophily, in particular when he states that *similarities need not lead to focused (clustered) interaction, and focused interaction can exist apart from similarity of individual characteristics* [Feld 1981].

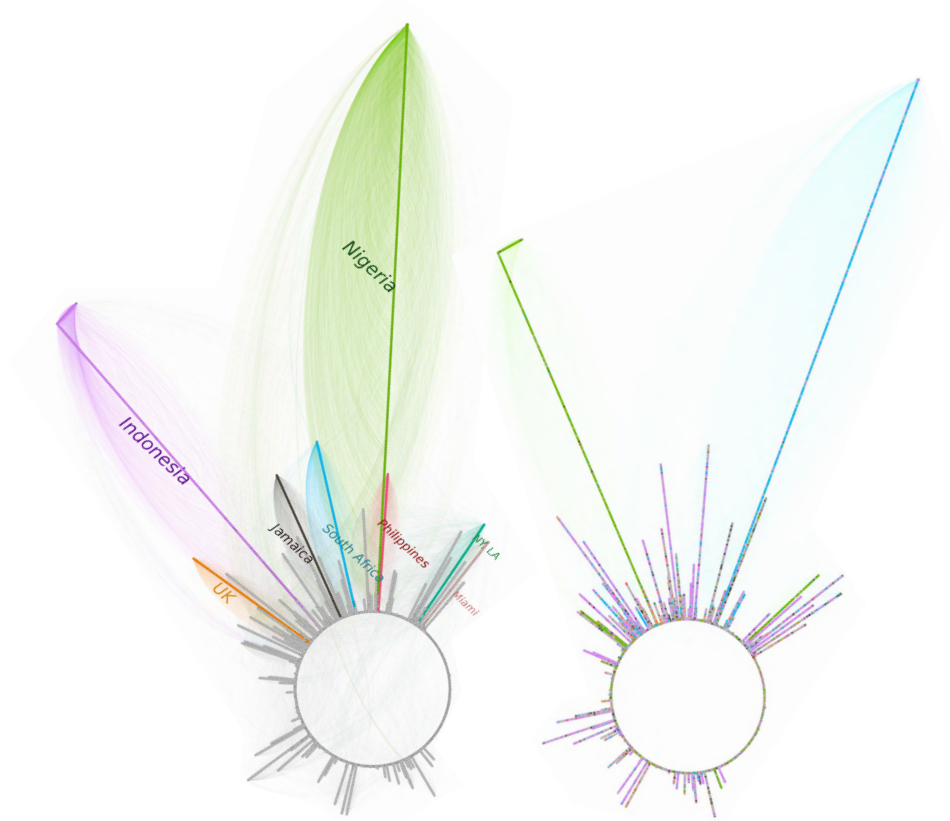


Fig. 15: **Modular** communication network **communities**; (left) radial axis visualization in Gephi [Bastian et al. 2009] of communication network communities with displayed identified user geolocation entities in each community; (right) SR communities produced by Infomap visualized with different colors on the communication network community representation; we can see to what extent the largest modular communities from the communication layer overlap with those produced from the semantic layer

Table VII: Community similarity between communication and semantic layer

<i>P communities</i>	<i>L communities</i>	<i>S</i>
P_0 - Philippines	L_{326}	0.41
P_8 - Nigeria	L_2	0.45
P_{10} - Indonesia	L_{159}	0.18
P_{11} - Nigeria	L_2	0.18
P_{102} - UK	L_{211}	0.13

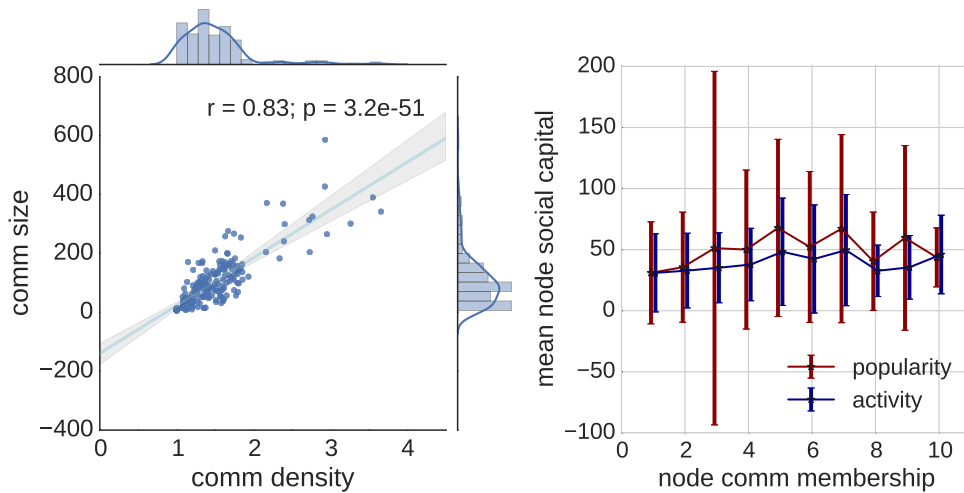


Fig. 16: **Overlapping communities**: (left) relationship between community size and density; (right) forms of social capital as a function of node community membership. We average values for 10 and more community memberships, due to data scarcity. Error bars show one standard deviation.

6.2. Overlapping communication communities

Next we analyze overlapping structural communities in communication network. We select the algorithm BigClam [Yang and Leskovec 2013] because it detects overlapping communities as the groups of nodes with denser links presence, in agreement with sociological theories, such as the Feld's [Yang and Leskovec 2014]. BigClam automatically detects 198 communities in our network, largest in size consisting of 586 users. **Community membership** of a user (defined as the number of communities in which it belongs), ranges from the minimum 1, for a majority of users, to the maximum 14, for a small number of users, and it exponentially decreases. Similar semantic analysis as with modular communities reveals that geographic foci are again the strongest predictor of communities. The subtle difference, however, is seen in modular communities being broken apart in several overlapping communities. For instance, the largest Nigerian modular community now has 7 overlapping counterparts. Many of the nodes from one modular community will belong to several such counterparts. By careful analysis, we reveal other foci, behind the overarching geographic, that drive such overlapping communities within the modular (these foci can again be geographic or not). For example, within the Nigerian group, we find subgroups discussing different geo-entities, in addition to common Nigeria: some talk about Ghana, some about Zambia and others about London. That not only geographic foci drive these overlapping sub-communities, we can see from the case for Malaysia where one subcommunity of 260 users has the predominant entity *selamat hari raya*, or Muslim greeting for Happy Eid. We also find communities around specialized topics, such as one of 144 users talking predominantly about NASCAR (auto racing). Hence, *our semantic analysis of overlapping community structure reveals that geographic and language foci are the largest foci, in terms of number of users connected. Within these foci as enablers, we can find other more focused and overlapping foci, with smaller number of users discussing more specific topics.*

Additionally, we look into which communities are featuring most overlaps with others. To this purpose, we introduce **community density** as the average number of community memberships for the nodes in the community. As presented in Fig. 16 (left), there is a strong positive correlation between the size of the community and its introduced density. Such result exhibits that the largest communities are those that feature most overlaps with other (sub)communities. Thinking of foci, such result can be interpreted also in the following way. The largest foci are as well enablers for

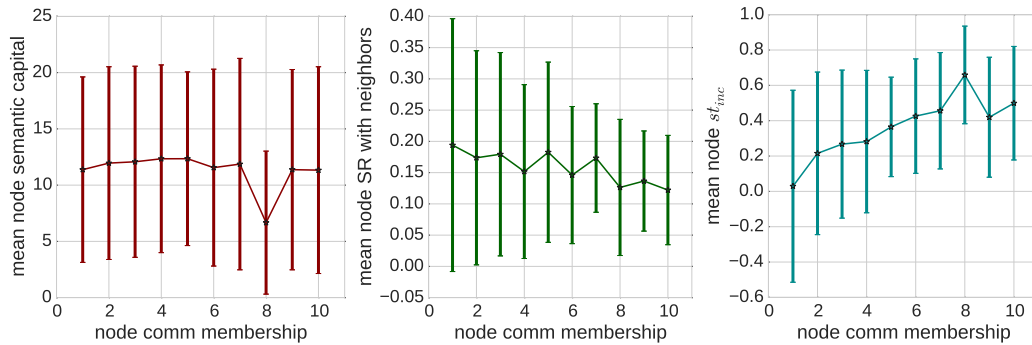


Fig. 17: **Pluralistic homophily and semantic capital:** (left) semantic capital, (middle) mean neighborhood SR and (right) status inconsistency in function of community memberships. Error bars show one standard deviation.

participating users to develop more additional foci of homophily. A related result in an analysis on Twitter is reported by Halberstam et al. [Halberstam and Knight 2014] who found that users *affiliated with majority political groups, relative to the minority group, have more connections, and are more densely connected.*

6.3. Pluralistic homophily

Pluralistic homophily results from several different foci. The users that share more communities (they are found in overlapping parts) have more homophilous foci in common, i.e., they feature aspects of pluralistic homophily. Such users are more densely connected [Yang and Leskovec 2013] forming a network core [Yang and Leskovec 2014]. Hence we ask whether the nodes in these parts tend to have higher *social capital*. However, we find no correlation between community membership and social capital, except that the nodes in more communities tend to be slightly more popular than active (see Fig. 16, right). Interestingly, the same holds for *semantic capital*: nodes with higher community membership are no more likely to be semantically rich than those belonging to less communities (see Fig. 17, left). This is a particularly surprising result, as such nodes, with higher community membership, are tied with their friends from different communities around different foci, according to the theory of focused interaction [Feld 1981]. We would expect them to be semantically richer, since they talk on several additional topics, as their community membership grows. However, as they are not semantically richer, then our next assumption is that such nodes must be less similar to their neighbors on average. Indeed, this is true, as presented in Fig. 17 (middle). The correlation between similarity to an average neighbor and community membership is highly negative and significant $-0.83, p = 0.003$. A concept of **opinion leaders** [Rogers and Bhowmik 1970] defines them as *the members of the group sought by others for opinion and advice and they are said to possess features and conformity to the norms that make them super-representative or similar to their average follower*. Hence, the users with increased community membership in our network potentially represent opinion leader within their different communities. So far we find no correlation between social or semantic capital and community membership. However, we ask what about *status inconsistency*. Since status inconsistency can be negative, we correlate its median value against community membership (although, similar result holds for mean value, as well). As presented in Fig. 17 (right), status inconsistency grows with community membership ($r = 0.87$ and $p = 0.001$). With this we reveal that the users in more communities do not have higher social or semantic status, but they can be characterized by increased status inconsistency. As mentioned in introduction, status inconsistency is suggested to be an attribute of individuals who are **drivers of social change** [Lanski

1954]. Therefore, we conclude that individuals featuring pluralistic homophily in communication networks are likely to be the opinion leaders and drivers of social change within their communities.

7. DISCUSSION AND CONCLUSION

Despite the vast and growing literature and research on what interweaves people in social networks, the interplay of homophily and influence as the main factors for social correlation with the network is still not fully explored and understood. Our first set of findings quantify to what extent semantic homophily and social influence affect the communication, its propensity and intensity in online social networks, though we are not trying to distinguish between these two factors. Concretely, we analyze interplay of semantic relatedness and communication intensity and show that while their correlation is low, their relationship is strongly captured by subtle communication network properties.

Next we show that several types of homophily are present in communication network, such as value (topics, sentiment) and status (social and semantic capital) homophily. Introduced social and semantic status metrics allow us to exhibit their growth with strength of the links (both, in terms of reciprocal communication and with increase in intensity). Assessment on how the two types of capital are affecting each other in communication network reveals diversity of relationships depending on which exact form of the two types of capitals is considered. While popularity and semantic capital are positively correlated, sentiment, inversely, is negatively correlated with social capital. In any case, we exhibit large diversity among users on the existing combinations of capitals they possess. Additional investigation on sociological concept of relative status reveals strong preference for communication with users of similar status. However, for relative social status particularly, we notice 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. Moreover, our data suggest a new hypothesis: this proposition holds only when both users are higher on the same status type, otherwise, communication intensity decreases compared to average.

Using temporal communication network we show that the tendencies of homophily and influence are dynamic and change their role and magnitude in time. In addition to confirming previous finding in other types of networks that similarity of users sharply grows before their link formation, we also explain in part the following decrease in similarity – as a result of link decommission. 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.

We analyze modular and overlapping community structure of the communication layer and find evidence for Feld's theory about focused organization of social ties. Comparison of best matching between community structure in communication and in semantic layer shows that cohesive communities cannot be explained only by semantic relatedness of users, instead there need to be a foci of homophily present around which communities are formed. Further analyses reveal that geographic foci are the largest predictor for both modular and overlapping communities. However, in the case of overlapping community structure, we find that such large foci also give space for smaller but stronger foci around which sub-communities within are formed. Precisely, larger foci tend to create denser communities (i.e., those with more overlapping parts within). Explanation from sociology is a tendency of people who are connected around one foci to find or create new foci to strengthen the interaction.

Finally, we also exhibit that pluralistic homophily does not correlate with social or semantic capital; instead the users who are connected with others around several different foci tend to have lower average similarity to those neighbors, while at the same time being increasingly status inconsistent.

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

Sociology; theory and questions	Experimental evidence	Novel hypotheses/open questions
<i>Quantification</i>		
Semantic homophily	Comm. propensity (c_p) and intensity (CI) increase with SR.	CI increases with status inconsistency , when both users are high on the same status dimension; otherwise CI decreases.
Status level homophily	(Un)directed degree assortativity; increases with tie strength and CI. \implies On higher CI, Twitter is more a social than information network.	in-in and out-out deg. assortativity coefficients obey a different pattern to others with increase in CI. Tendency of users with a particular popularity difference to interact.
Value level homophily	Attribute assortativity on semantic aspects of comm., such as topics, sentiment, semantic diversity.	Semantic diversity and negative sentiment increase with comm. activity.
<i>Temporal evolution</i>		
Semantic homophily evolution	Average increase through time in SR among communicating users. The increase is driven by semantic homophily and social influence.	Average increase through time in SR among users who never communicated . The increase is driven by external influence.
Heterophilous links dissolution	Dissolution more due to social status and less due to semantic value heterophily. Persisting pairs having more weak contacts are increasingly likely to stop communicating.	At the time of a link dissolution, one or both of the participating users are likely to have found a new contact that will replace the one being disconnected.
<i>Community foci</i>		
Theory of focused interaction	Semantic similarity in terms of SR only moderately explains structural communities. Modular communities explained by geolocation entities as comm. foci.	
Pluralistic homophily	Overlapping communities formed around other foci enabled by overarching geolocation foci.	High correlation between size of a community and its density of overlap. Pluralistic homophily is not explained by social or semantic capital. On the other hand, individuals exhibiting pluralistic homophily are increasingly status inconsistent .

7.1. Limitations

A limitation of our work posed by the restricted dataset is that we are not considering the entire Twitter channel for information flow, as there are also considerable amount of information flowing along the retweet network, which is not taken into consideration in this work. Besides this, the mention mechanism in Twitter can be sometimes biased towards specific target audiences for specific information [Tang et al. 2015].

Another limitation is that our results are solely about computer-mediated communication and we do not tackle the impact of Internet (online medium) on social interaction.

Further investigation is needed on the influence of the threshold for semantic relatedness on the semantic homophily, as we show in this work that the semantic layer became disassortative after threshold equal to 0.6. Additional and improved sentiment analysis is needed to understand how the social reinforcement influences communication between users and if there exists happiness paradox while people communicate in social network.

8. METHODS

In this section we describe how we build Wikipedia-based semantic database using an English pages dump (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* [Giuseppe Attardi 2015] to pre-process and clean the texts. The ESA algorithm is based on the TF-IDF (term frequency - inverse document frequency) [Baeza-Yates et al. 1999] scores of words in different articles in the Wikipedia corpus. As a result a word w_1 is mapped to the *concept vector* $CV(w_1) = \{(C_1^1, V_1^1), (C_2^1, V_2^1), (C_3^1, V_3^1), \dots, (C_{M_1}^1, V_{M_1}^1)\}$. C_j^1 represent Wikipedia concepts and V_j^1 are TF-IDF scores for the word w_1 in those articles and are calculated as follows:

$$V_j^1 = TF \cdot IDF = (1 + \log(f_{1,j})) \cdot \log\left(\frac{N}{n_t}\right), \quad (1)$$

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 [Pedregosa et al. 2011] 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 [Gabrilovich and Markovitch 2009] 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 is open-source and published on Github [Scepanovic 2016].

8.0.1. Word Semantic Relatedness. The semantic relatedness (SR) between words is not measured directly, but it is rather determined through a set of concepts highly related to them [Gabrilovich and Markovitch 2009; Hieu et al. 2013]. Let us assume that the SR between words w_1 and w_2 is requested. The word SR 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) of Wikipedia which are related to the words. Let us assume that w_1 is mapped to *concept (tf-idf) vector*: $CV(w_1) = \{(C_1^1, V_1^1), (C_2^1, V_2^1), (C_3^1, V_3^1), \dots, (C_{M_1}^1, V_{M_1}^1)\}$ and w_2 is mapped to *concept (tf-idf) vector*: $CV(w_2) = \{(C_1^2, V_1^2), (C_2^2, V_2^2), (C_3^2, V_3^2), \dots, (C_{M_2}^2, V_{M_2}^2)\}$. These are the sets of Wikipedia concepts, C_j^1 and C_j^2 , which are related to the word w_1 and w_2 and their TF-IDF scores, V_j^1 and V_j^2 , respectively. In the following, we will assume that N is the number of common concepts in $CV(w_1)$ and $CV(w_2)$.
- **Calculating the SR between words using cosine similarity between obtained CVs.** For measuring the degree of semantic relatedness, cosine similarity between the CVs for two words w_1 and w_2 is calculated. This measure gives the cosine of the angle between the two vectors $CV(w_1)$

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

$$SR(w_1, w_2) = \cos(CV(w_1), CV(w_2)) = \frac{\sum_{i=1}^N V_i^1 \cdot V_i^2}{\sqrt{\sum_{k=1}^{M_1} (V_k^1)^2} \cdot \sqrt{\sum_{l=1}^{M_2} (V_l^2)^2}}, \quad (2)$$

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) as the TF-IDF weights can not be negative.

8.0.2. Document Semantic Relatedness. The semantic relatedness (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 document SR calculation follows the three steps below.

- **Analyzing documents using the term frequency (TF) approach which finds the frequency of words in the document.** The result of this step is a list of important words with their corresponding TF scores. Let us assume that:
 d_1 is analyzed to *term (tf) vector*: $T(d_1) = \{(t_1^1, v_1^1), (t_2^1, v_2^1), (t_3^1, v_3^1), \dots, (t_m^1, v_m^1)\}$,
 d_2 to *term (tf) vector*: $T(d_2) = \{(t_1^2, v_1^2), (t_2^2, v_2^2), (t_3^2, v_3^2), \dots, (t_n^2, v_n^2)\}$, 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 8.0.1). For instance, the t_1^1 term is mapped to *concept (tf-idf) vector*: $CV(t_1^1) = \{(C_1^1, (v_1^1 \times V_1^1)), (C_2^1, (v_1^1 \times V_2^1)), (C_3^1, (v_1^1 \times V_3^1)), \dots, (C_M^1, (v_1^1 \times V_M^1))\}$. The other terms in $T(d_1)$ can be represented in a similar way. 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 the list of relevant Wikipedia concepts and TF-IDF scores is found in $CV(d_2)$.
- **Calculating the SR between documents using 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. 2).

8.0.3. 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 [Gabrilovich and Markovitch 2009], 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 [Faruqui and Dyer 2014]. The results of the evaluation are presented in Table IX. 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 [Cramer 2008]. However, our evaluation scores are comparable to the original implementation [Gabrilovich and Markovitch 2009] and to the ESA implementations available online.

ACKNOWLEDGMENTS

S.Š. research was partially financed by CIVIS EU FP7 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

Table IX: SR knowledge database evaluation

<i>Human judgments dataset</i>	<i>Spearman's rank</i>	<i>Pearson's correlation</i>
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

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. 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

- Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. 2012. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)* 6, 2 (2012), 9.
- An IBM Company. 2016. AlchemyAPI and IBM Watson. (January 2016). <http://www.alchemyapi.com/api>
- Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. 2008. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 7–15.
- Sinan Aral, Lev Muchnik, and Arun Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* 106, 51 (2009), 21544–21549.
- Sinan Aral and Dylan Walker. 2012. Identifying influential and susceptible members of social networks. *Science* 337, 6092 (2012), 337–341.
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, and others. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.
- Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 519–528.
- George A Barnett and Grace A Benefield. 2015. Predicting international Facebook ties through cultural homophily and other factors. *New Media & Society* (2015), 1461444815604421.
- Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. (2009). <http://www.aiai.org/ocs/index.php/ICWSM/09/paper/view/154>
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. " O'Reilly Media, Inc."
- Catherine A Bliss, Isabel M Kloumann, Kameron Decker Harris, Christopher M Danforth, and Peter Sheridan Dodds. 2012. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science* 3, 5 (2012), 388–397.
- Per Block and Thomas Grund. 2014. Multidimensional homophily in friendship networks. *Network Science* 2, 02 (2014), 189–212.

- Vincent Blondel, Gautier Krings, Isabelle Thomas, and others. 2010. Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies* (2010).
- Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.
- Johan Bollen, Bruno Gonçalves, Guangchen Ruan, and Huina Mao. 2011. Happiness is assortative in online social networks. *Artificial life* 17, 3 (2011), 237–251.
- Pierre Bourdieu. 2011. The forms of capital.(1986). *Cultural theory: An anthology* (2011), 81–93.
- Ronald S Burt. 2000. Decay functions. *Social networks* 22, 1 (2000), 1–28.
- Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. *ICWSM* 133 (2011), 89–96.
- Irene Cramer. 2008. 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, 59–70.
- David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. 2008. 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, 160–168.
- Munmun De Choudhury. 2011. Tie Formation on Twitter: Homophily and Structure of Egocentric Networks. *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing* (Oct. 2011), 465–470. DOI :<http://dx.doi.org/10.1109/PASSAT/SocialCom.2011.177>
- Munmun De Choudhury, Hari Sundaram, Ajita John, Doree Duncan Seligmann, and Aisling Kelliher. 2010. " Birds of a Feather": Does User Homophily Impact Information Diffusion in Social Media? *arXiv preprint arXiv:1006.1702* (2010).
- Young-Ho Eom and Hang-Hyun Jo. 2014. Generalized friendship paradox in complex networks: The case of scientific collaboration. *Scientific reports* 4 (2014).
- Ernesto Estrada. 2011. Combinatorial study of degree assortativity in networks. *Physical Review E* 84, 4 (2011), 047101.
- Manaal Faruqui and Chris Dyer. 2014. 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.
- Scott L Feld. 1981. The focused organization of social ties. *American journal of sociology* (1981), 1015–1035.
- Diane Felmlee, Susan Sprecher, and Edward Bassin. 1990. The dissolution of intimate relationships: A hazard model. *Social Psychology Quarterly* (1990), 13–30.
- Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis.. In *IJCAI*, Vol. 7. 1606–1611.
- Evgeniy Gabrilovich and Shaul Markovitch. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research* (2009), 443–498.
- Giuseppe Attardi. 2015. Wikipedia Extractor. (April 2015). <https://github.com/attardi/wikiextractor>
- Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology* 78, 6 (1973), 1360–1380.
- Yosh Halberstam and Brian Knight. 2014. *Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter*. Technical Report. National Bureau of Economic Research.

- Sebastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. 2015. Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies* 8, 1 (2015), 1–254.
- Nguyen Trung Hieu, Mario Di Francesco, and Antti Ylä-Jääski. 2013. Extracting knowledge from wikipedia articles through distributed semantic analysis. In *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies*. ACM, 6.
- Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*.
- Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and P Krishna Gummadi. 2012. Geographic Dissection of the Twitter Network.. In *ICWSM*.
- Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. ACM, 591–600.
- Timothy La Fond and Jennifer Neville. 2010. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th international conference on World wide web*. ACM, 601–610.
- Paul F Lazarsfeld and Robert K Merton. 1954. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society* 18 (1954), 18–66.
- RTAJ Leenders. 1997. Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection. *Evolution of social networks* 1 (1997).
- Adrienne Lehrer and Keith Lehrer. 1982. Antonymy. *Linguistics and philosophy* 5, 4 (1982), 483–501.
- Gerhard E Lenski. 1954. Status crystallization: a non-vertical dimension of social status. *American sociological review* 19, 4 (1954), 405–413.
- Jure Leskovec and Eric Horvitz. 2008. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 915–924.
- Clement Levallois. 2013. Umigon: sentiment analysis for tweets based on terms lists and heuristics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, Vol. 2. 414–417.
- James R Lincoln and Jon Miller. 1979. Work and friendship ties in organizations: A comparative analysis of relation networks. *Administrative science quarterly* (1979), 181–199.
- Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.
- M McPherson, L Smith-Lovin, and JM Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 2001 (2001), 415–444. <http://www.jstor.org/stable/10.2307/2678628>
- Kevin Meehan, Tom Lunney, Kevin Curran, and Aiden McCaughey. 2013. Context-aware intelligent recommendation system for tourism. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*. IEEE, 328–331.
- Folke Mitzlaff, Martin Atzmueller, Andreas Hotho, and Gerd Stumme. 2014. The social distributional hypothesis: a pragmatic proxy for homophily in online social networks. *Social Network Analysis and Mining* 4, 1 (2014), 1–14.
- M Lynne Murphy. 2003. *Semantic relations and the lexicon: antonymy, synonymy and other paradigms*. Cambridge University Press.
- Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. 2014. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 493–498.

- Mark EJ Newman. 2003. Mixing patterns in networks. *Physical Review E* 67, 2 (2003), 026126.
- Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103, 23 (2006), 8577–8582.
- Mark EJ Newman and Juyong Park. 2003. Why social networks are different from other types of networks. *Physical Review E* 68, 3 (2003), 036122.
- M. E. J. Newman. 2002. Assortative Mixing in Networks. *Phys. Rev. Lett.* 89 (Oct 2002), 208701. Issue 20. DOI : <http://dx.doi.org/10.1103/PhysRevLett.89.208701>
- Hans Noel and Brendan Nyhan. 2011. The unfriending problem: The consequences of homophily in friendship retention for causal estimates of social influence. *Social Networks* 33, 3 (2011), 211–218.
- 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, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- Mahendra Piraveenan, Mikhail Prokopenko, and Albert Zomaya. 2012. Assortative mixing in directed biological networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 9, 1 (2012), 66–78.
- Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. Do all birds tweet the same?: characterizing twitter around the world. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 1025–1030.
- Alejandro Portes. 2000. Social capital: Its origins and applications in modern sociology. *LESSER, Eric L. Knowledge and Social Capital*. Boston: Butterworth-Heinemann (2000), 43–67.
- Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. SentiBench—a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5, 1 (2016), 1–29.
- Giuseppe Rizzo and Raphaël Troncy. 2011. Nerd: evaluating named entity recognition tools in the web of data. (2011).
- Everett M Rogers and Dilip K Bhowmik. 1970. Homophily-heterophily: Relational concepts for communication research. *Public opinion quarterly* 34, 4 (1970), 523–538.
- Martin Rosvall and Carl T Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105, 4 (2008), 1118–1123.
- Camille Roth. 2005. Generalized preferential attachment: Towards realistic socio-semantic network models. In *ISWC 4th Intl Semantic Web Conference, Workshop on Semantic Network Analysis*, Vol. 171. 1613–0073.
- Camille Roth and Jean-Philippe Cointet. 2010. Social and semantic coevolution in knowledge networks. *Social Networks* 32, 1 (2010), 16–29.
- Hassan Saif, Yulan He, and Harith Alani. 2012. Semantic sentiment analysis of twitter. In *International Semantic Web Conference*. Springer, 508–524.
- Sanja Scepanovic. 2016. Implementation of ESA algorithm for a Wikipedia SR database. (April 2016). DOI : <http://dx.doi.org/10.5281/zenodo.49750>
- Cosma Rohilla Shalizi and Andrew C Thomas. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research* 40, 2 (2011), 211–239.
- Shujing Sun and Huaxia Rui. 2017. Link Formation on Twitter: The Role of Achieved Status and Value Homophily. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

- Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. 2013. Exploiting homophily effect for trust prediction. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 53–62.
- Liyang Tang, Zhiwei Ni, Hui Xiong, and Hengshu Zhu. 2015. Locating targets through mention in Twitter. *World Wide Web* 18, 4 (2015), 1019–1049.
- Mike Thelwall. 2013. Heart and soul: Sentiment strength detection in the social web with sentistrength. *Proceedings of the CyberEmotions* (2013), 1–14.
- Crispin Thurlow, Laura Lengel, and Alice Tomic. 2004. *Computer mediated communication*. Sage.
- Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. 2011. The anatomy of the facebook social graph. *arXiv preprint arXiv: ...* (Nov. 2011), 17. <http://arxiv.org/abs/1111.4503>
- Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. 2009. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*. Acn, 205–218.
- Jaewon Yang and Jure Leskovec. 2012. Community-affiliation graph model for overlapping network community detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 1170–1175.
- Jaewon Yang and Jure Leskovec. 2013. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 587–596.
- Jaewon Yang and Jure Leskovec. 2014. Overlapping communities explain core–periphery organization of networks. *Proc. IEEE* 102, 12 (2014), 1892–1902.
- Jaewon Yang and Jure Leskovec. 2015. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* 42, 1 (2015), 181–213.
- Jaewon Yang, Julian McAuley, and Jure Leskovec. 2014. Detecting cohesive and 2-mode communities in directed and undirected networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 323–332.
- Mustafa Yavaş and Gönenç Yücel. 2014. Impact of homophily on diffusion dynamics over social networks. *Social Science Computer Review* (2014), 0894439313512464.
- Xiaohua Zeng and Liyuan Wei. 2013. Social ties and user content generation: Evidence from Flickr. *Information Systems Research* 24, 1 (2013), 71–87.