




Link transmission centrality in large-scale social networks

Qian Zhang^{1*} , Márton Karsai^{1,2} and Alessandro Vespignani¹ 

*Correspondence:

qianzhang@northeastern.edu

¹Laboratory for the Modeling Biological and Socio-technical Systems, Northeastern University, Boston, USA

Full list of author information is available at the end of the article

Abstract

Understanding the importance of links in transmitting information in a network can provide ways to hinder or postpone ongoing dynamical phenomena like the spreading of epidemic or the diffusion of information. In this work, we propose a new measure based on stochastic diffusion processes, the *transmission centrality*, that captures the importance of links by estimating the average number of nodes to whom they transfer information during a global spreading diffusion process. We propose a simple algorithmic solution to compute transmission centrality and to approximate it in very large networks at low computational cost. Finally we apply transmission centrality in the identification of weak ties in three large empirical social networks, showing that this metric outperforms other centrality measures in identifying links that drive spreading processes in a social network.

Keywords: Social networks; Link centrality measures; Diffusion processes; Weak tie

1 Introduction

The importance of nodes and links in networks is commonly measured through centrality measures. Their definitions generally rely on local and/or global structural information. Centrality measures using local information, like the node degree or link overlap, are computed efficiently as they only require knowledge about the neighbors of a given node or link. On the other hand, these measures cannot provide information on which nodes or links play global roles in the network structure. On the contrary, centrality measures based on global information about the network structure, like betweenness and closeness centrality [1, 2], Katz centrality [3], k-shell index [4, 5], subgraph centrality [6] and induced centrality measures [7] may better characterize the overall importance of a node or link. Unfortunately, although effective algorithms for approximating these quantities have recently been proposed [8, 9], estimating these measures in large scale networks is still computationally challenging.

While global centrality measures have been very successful in identifying structurally important nodes or links in networks, it has been argued [10] that they do not evidently identify nodes or links with a key role in dynamical processes. Other centrality metrics, which directly use dynamical processes to assign importance were found to be more successful in this sense. The best examples are metrics based on random walkers like Page-Rank [11], eigenvector centrality [12], or accessibility [13]. Other examples are local metrics like the expected force [14], or percolation centrality [15]. These measures are based

on random diffusion processes, but do not fully capture the basic mechanisms behind contagion mediated spreading phenomena. Here we define a new link centrality measure, *transmission centrality*, tailored to identify the role of nodes and links in controlling contagion phenomena. The transmission centrality measures the average number of nodes who are reached by the contagion process through each link during the spreading of a stochastic contagion process. This provides a direct measure of the centrality of the link in hindering or facilitating the contagion process. In real contagion processes, links correspond to specific interactions among individuals, or specific exchanges of information, or individuals in the case that the nodes represents specific subpopulations. Controlling single contagion routes instead of completely isolating an individual may result in a convenient option for mitigating the spreading of epidemics or enhancing the velocity of information diffusion [16–18]. In the case of very large-scale network, we propose a heuristic calculation of transmission centrality, which is both computationally efficient and can be easily extended for weighted, directed, or temporal networks or even for nodes. Furthermore, to demonstrate the usefulness of transmission centrality we present a case study where we use this metric to identify weak ties [19, 20] in social networks and characterize their role in contagion processes.

As it follows, after a brief discussion of related works and utilized datasets, we formally introduce transmission centrality and discuss a heuristic method for its approximate calculation. Then we discuss its properties and correlations with local centrality measures in three large-scale real world social networks. Finally, we present simulation results of SIR spreading processes to demonstrate the capacity of combined local measures and transmission centrality in designing effective strategies to enhance or hinder information diffusion in social networks.

2 Related works

Node centralities have been widely studied, from classical static centralities like degree, closeness, betweenness, eigenvector [21] to centrality measures based on dynamical processes, such as random walk (e.g. PageRank [11]). Among these, betweenness centrality is one of the most popular measures as it quantifies the importance of a node by considering the global structure of a network instead of local information. Unfortunately, the efficiency of algorithms to calculate betweenness centrality is still challenging in the case of large-scale social networks as its best computation method has $\mathcal{O}(|V||E|)$ complexity for unweighted networks and $\mathcal{O}(|V||E| + |V|^2 \log |V|)$ for weighted networks [8]. While many variants and approximation algorithms have been proposed to improve its algorithmic efficiency [22–27], researchers have also proposed alternative measures to quantify the importance of nodes in terms of dynamical processes on top of a network, such as K-path centrality [28] and percolation centrality [15]. K-path centrality [28] applies self-avoiding random walks of length k and counts the probability that a message originating from a given source traverses a node. The percolation centrality [15] measures the relative importance of a node based on both network structure and its percolated states. Single-node-influence centrality and Shapley centrality assess the importance of a node in isolation and in a group respectively in social influence propagation processes [29]. [30] simulates epidemic models (SIS and SIR) to estimate node centralities on top of temporal social networks. Interestingly, this study shows that spreading processes fail to characterize the centrality measures like degree and core numbers of infected nodes. Dynamics-sensitive centrality [31], which counts the outbreak size in an epidemic model to quantify

spreading influence of nodes, can better capture the importance of nodes particularly in epidemic spreading processes.

Most centrality algorithms have also been generalized to the estimation of link centrality measures, such as edge betweenness centrality, spanning edge betweenness centrality [32, 33], and K-path edge centrality [34]. As node centralities aim to characterize the importance of nodes in a network, edge centralities provide quantitative perspectives to measure the importance of links in a network structure [35–40].

3 Materials and methods

3.1 Network data descriptions

In the following study, we will discuss centrality algorithms by using three distinct sets of data recording communications between thousands or millions of individuals. For each dataset, first we aggregate the sequence of interactions to a static social network, excluding possible commercial communications. To do so, we only draw links between individuals who had at least one pair of mutual interactions during the observation period. In addition, to avoid leaf links we extract the k -core ($k = 2$) structure [41, 42] of each network and use their largest connected component (LCC).

The first dataset we investigate is collected from the mobile phone call (MPC) communication sequences of 4,256,137 individuals during 4 weeks with 1 second resolution [43, 44]. Individuals are anonymous users of a single operator with 20% market share in a European country. The static social network contains 5,279,169 mutual links. The final k -core ($k = 2$) structure of the LCC includes 1,926,787 nodes and 3,269,634 edges.

The second social network is aggregated from the sequence of wall posts of Facebook users (FB) [45–47]. The data records interactions from September 2004 to January 2009 between 31,720 users connected by 80,592 mutual links. The k -core ($k = 2$) structure of the LCC of this network contains 20,244 nodes and 70,132 edges.

The last social network is a Twitter conversation network (TW), which was constructed from tweets from October 2010 to November 2013, which were collected through the Twitter Gardenhose [48]. We restrict our dataset to tweets with live GPS coordinates providing us over 420 million communication events, which represent a 1–2% of the entire volume. We construct a social network based on mutual conversational tweets (*@mentions*) between 4,155,700 users connected by 6,506,519 links. The k -core ($k = 2$) structure of the LCC of the Twitter conversation network consists of 966,779 nodes linked by 2,779,524 edges.

3.2 Transmission centrality

Transmission centrality aims to measure for each link in a network its influence in disseminating some globally spreading information. More precisely it measures the number of nodes who received information during a diffusion process through a given link. Its definition intrinsically assumes a diffusion process to unfold on a network structure. In our definition we use the simplest possible information spreading process, the *Susceptible-Infected model* [49], however this can be replaced by any other diffusion process. The Susceptible-Infected (SI) process is defined on a connected network $G = (V, E)$, where nodes $u \in G.V$ can be in two mutually exclusive states, either susceptible (S) or infected (I). Initially each node is susceptible (S) except a randomly selected seed node, which is set to be in state I . In one iteration step each infected node can infect its susceptible neighbors with rate β until

Input: $G = (V, E)$, β , s

Output: $G_{BT} = (V_{BT}, E_{BT})$ *branching tree of spreading*

```

1:  $Q = \text{queue}()$  // queue of  $I$  nodes with susceptible neighbors
2:  $G_{BT}.V_{BT} = \emptyset$  // the branching tree
3:  $G_{BT}.E_{BT} = \emptyset$ 
4: for each vertex  $u \in G.V - \{s\}$  do
5:    $u.state = S$ 
6:    $u.asc = NIL$ 
7: end for
8:  $s.state = I$ 
9: ENQUEUE( $Q, s$ )
10: while  $Q \neq \emptyset$  do
11:    $u = \text{DEQUEUE}(Q)$ 
12:   SN = False // remaining susceptible neighbor of node  $u$ 
13:   for each  $v \in \text{sort}(G.Adj[u])$  do // we check neighbors of node  $u$  in a sorted
      fashion
14:     if ( $v.state == S$ ) then
15:       if ( $\text{rand}() \leq \beta$ ) then
16:          $v.state = I$ 
17:          $u.asc = u$ 
18:          $G_{BT}.V_{BT}.add(v)$ 
19:          $G_{BT}.E_{BT}.add((u, v))$ 
20:         ENQUEUE( $Q, v$ )
21:       else
22:         SN = True
23:       end if
24:     end if
25:     if SN == True then
26:       ENQUEUE( $Q, u$ )
27:     end if
28:   end for
29: end while

```

Algorithm 1: Susceptible-infected process

every node becomes infected in the network. Note that the parameter β here scales with the speed of information spreading, with value $\beta = 1$ corresponding to the fastest possible information diffusion process determining the shortest diffusion routes between the seed and any other node in the network. (We set $\beta = 1$ in this study if not noted otherwise.) This diffusion process can be simulated with a modified breath-first-search algorithm [50] as written in Alg. 1. There, during the unfolding of the diffusion we keep infected nodes with susceptible neighbors in a Q queue and record the branching tree $G_{BT} = (V_{BT}, E_{BT})$ of the process by keeping track of the direct ascendant of each node from which it received the information. Note that by exploring the neighbors of an infected node in a sorted fashion (see line 13 in Alg. 1) makes this algorithm fully deterministic in case of $\beta = 1$. Exploiting

the structure of the actual branching tree, *transmission centrality* is formally defined as

$$C_{tr}(u, v) = \begin{cases} \max(|\text{desc}(u)|, |\text{desc}(v)|), & \text{if } (u, v) \in E_{BT}, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $|\text{desc}(i)|$ denotes the number of descendant nodes of node i in the branching tree of the actual spreading.

The branching tree G_{BT} , which is a subgraph of G , encodes the diffusion paths that the information takes to reach the vertices of the network. Using its structure we can easily measure the actual C_{tr} value of each link by performing a second step of calculation based on the *river-basin algorithm* [51]. In practice, taking the initial seed s as the root of G_{BT} , and starting from the leaves of the branching tree we can count the number of descendant nodes of each link, i.e., who received the information via the actual link. The algorithm is summarized in Alg. 2, illustrated in Fig. 1.

Input: $G = (V, E)$ and $G_{BT} = (V_{BT}, E_{BT})$

Output: C_{tr} dictionary of *transmission centrality* values

```

1:  $C_{tr} = dict()$ 
2: for  $(u, v) \in G.E$  do
3:    $C_{tr}((u, v)) = 0$  // set counter to zero for each link
4: end for
5: while  $G_{BT}.E_{BT} \neq \emptyset$  do
6:   for  $v \in G_{BT}.V_{BT}$  do
7:     if  $k_v == 1$  then
8:        $p = asc(v)$  // parent node of  $v$ 
9:        $gp = asc(p)$  // grandparent node of  $v$ 
10:       $C_{tr}((v, p)) = C_{tr}((v, p)) + 1$ 
11:       $C_{tr}((p, gp)) = C_{tr}((p, gp)) + C_{tr}((v, p))$ 
12:       $G_{BT}.E_{BT} \leftarrow G_{BT}.E_{BT} - \{(v, p)\}$ 
13:       $G_{BT}.V_{BT} \leftarrow G_{BT}.V_{BT} - \{v\}$ 
14:     end if
15:   end for
16: end while

```

Algorithm 2: Transmission centrality

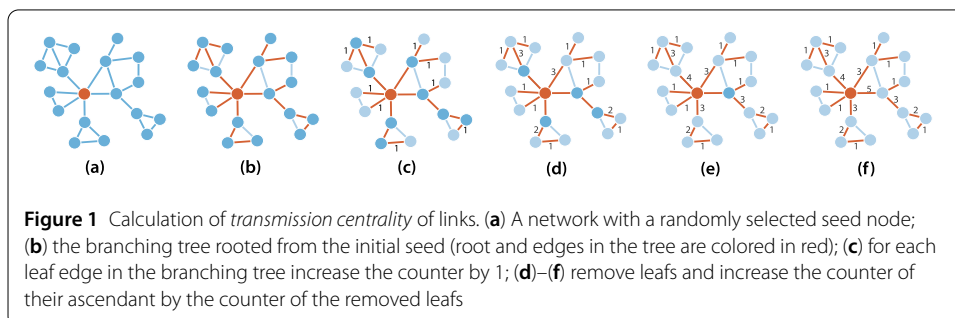


Figure 1 Calculation of *transmission centrality* of links. (a) A network with a randomly selected seed node; (b) the branching tree rooted from the initial seed (root and edges in the tree are colored in red); (c) for each leaf edge in the branching tree increase the counter by 1; (d)–(f) remove leaves and increase the counter of their ascendant by the counter of the removed leaves

Input: $G = (V, E), \beta$
Output: C_{tr}^{avr} dictionary of *average transmission centrality* values

```

1:  $C_{tr}^{avr} = dict()$ 
2: for  $(u, v) \in G.E$  do
3:    $C_{tr}^{avr}((u, v)) = 0$  // set counter to zero for each link
4: end for
5: for  $v \in G.V$  do
6:    $G_{BT} \leftarrow SusceptibleInfected(G, \beta)$ 
7:    $C_{tr}^{act} \leftarrow TransmissionCentrality(G, G_{BT})$ 
8:   for  $(u, v) \in G.E$  do
9:      $C_{tr}^{avr}((u, v))_+ = C_{tr}^{act}((u, v))$  // summing realisations
10:  end for
11: end for
12: for  $(u, v) \in G.E$  do
13:    $C_{tr}^{avr}((u, v)) = C_{tr}^{avr}((u, v)) / |G.V|$  // computing averages
14: end for

```

Algorithm 3: Average transmission centrality

First we define a dictionary C_{tr} , which associates a counter to each link $(i, j) \in G.E$, that we set to zero initially (lines 1–3 in Alg. 2). Then we recursively do the following for every node $v \in G_{BT}.V_{BT}$, which appears with degree $k_v = 1$ in G_{BT} :

- (a) Increase by one the counter $C_{tr}((v, p))$ of the (leaf) edge $e_f = (v, p) \in G_{BT}.E_{BT}$, which connects v to its parent node $p = asc_{BT}(v)$ in $G_{BT}.V_{BT}$ (line 10 in Alg. 2).
- (b) Increase by $C_{tr}((v, p))$ the counter $C_{tr}((p, gp))$ of its ascendant edge $asc_{BT}(e_f) = (p, gp)$, where $gp = asc(p)$ is the grandparent node of v in $G_{BT}.V_{BT}$ (line 11 in Alg. 2).
- (c) Remove v from $G_{BT}.V_{BT}$ and e_f from $G_{BT}.E_{BT}$ (lines 12 and 13 in Alg. 2). The final transmission centrality value of the actual link $e_f = (v, p)$ is stored in $C((v, p))$.

By repeating II.(a)–(c) recursively for each appearing leaf edge we assign a non-zero value for each link in the branching tree as it is demonstrated in Fig. 1(c)–(f).

The transmission centrality of a link can take values between 0 (for links, which are not in the branching tree) and $(N - 1)$ (e.g. in the case the seed is propagating information via a single link). Its actual value depends on the choice of the seed node and on the structure of the branching tree determined by the stochastic diffusion process. In this way centrality values of the same link may vary from one realization to another. To eliminate the effects of such fluctuations the final definition of transmission centrality of links is taken as the average centrality value for each link computed over processes initiated from every node in the network (for a algorithmic definition see Alg. 3). Note that from now on C_{tr} always assigns an average quantity if not stated otherwise.

4 Results

4.1 Heuristic calculation of transmission centrality

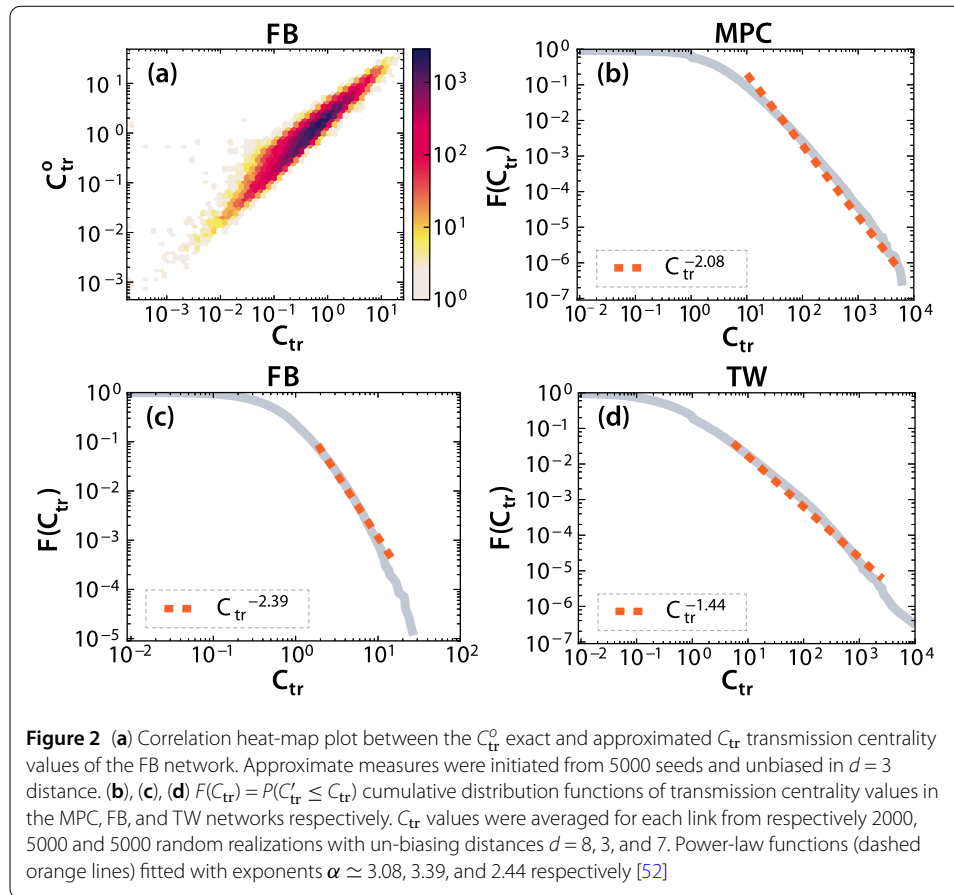
One iteration to measure C_{tr} performs with $\mathcal{O}(|E|)$ time complexity, in this case where we initiate its calculation from every node $v \in V$, its overall complexity is $\mathcal{O}(|V||E|)$. It is however possible to define a heuristic estimate of transmission centrality at a considerably small computational cost. As the branching trees of different realizations may largely

overlap, a relatively small number of independent realizations, initiated from a reduced set of randomly selected seeds, could provide a good approximation to transmission centrality. Link transmission centrality initiated from a single node provides a locally biased measure as it assigns higher values to links closer to the actual seed. This bias is averaged out if we initiate the spreading process from every node in the network, but in case of a limited number of seeds it has residual effects. One way to eliminate this residual bias is by assigning zero centrality values to links connecting nodes closer than a distance d to the actual seed. The best value of d depends on the network; however this can be estimated by parameter scanning, as demonstrated in Fig. S1 (Additional file 1).

To illustrate the computation of the heuristic estimate, we use the FB network with 20,244 nodes (for more details see Sect. 3.1) and compute the average transmission centrality for each link via the exact method by initiating an SI process from each node and the heuristic method where we initiate processes from 5000 random seeds (i.e. $\sim 25\%$ of all nodes) and eliminate biases in distance $d = 3$ around each seed (for more on the selection of this value see Fig. S1 (Additional file 1)). In Fig. 2(a) we present a heat-map plot about the correlation between the exact (assigned as C_{tr}^e here) and the approximated (assigned as C_{tr}) centrality values of each link. It is evident that there is a strong correlation between the exact and approximated values of centralities, quantified by an $r = 0.96$ ($p < 10^{-6}$) Pearson correlation coefficient. Consequently, this unbiased sampling method can provide very close approximations to the exact transmission centrality values, while considerably reducing the computational cost ($\sim 25\%$ in this case). Note that this correlation analysis was not repeated for the other two empirical networks as the computation of the exact method would take extremely long on such large networks due to its computational complexity.

Subsequently, we applied the approximate method to compute transmission centrality in the MPC network (with 2000 seeds and $d = 8$) and TW network (with 5000 seeds and $d = 7$) as well. We consistently found that the average unbiased transmission centrality of links, measured in the three empirical systems, are broadly distributed (see in Fig. 2(b)–(d) respectively for the MPC, FB and TW networks) with power-law tails with exponents $\alpha = 3.08, 3.39$ and 2.44 for the MPC, FB and TW networks respectively, determined by the fitting method explained in [52]. This demonstrates the high variance of importance of links in transmitting information, which can be duly the consequence of the community rich structure of the three investigated social networks.

Transmission centrality can be generalized in various ways. First, it can be easily defined as a *node centrality metric* by counting for each node the number of their descendant nodes in the branching tree. Moreover it can be extended for *directed and/or weighted networks* by restricting the SI process to respect the direction of links during spreading or by scaling the transmission rate with the normalized weight of links. In addition, for an SI process one can explore central links in the case when the process does not diffuse along the shortest paths. By taking $\beta < 1$, longer spreading paths become plausible allowing the inference of links, which are central in any scenario. Transmission centrality can be easily defined for *temporal networks* [53] as well. Contrary to static networks, in temporal structures information can transmit between nodes only at the time of their interactions. As a result, information can travel only along time-respecting paths in the network, which drastically restricts the final outcome of any global contagion processes [54] and has evident consequences on the measured centrality values. Links, which appeared unimportant in the



static structure may be central in the temporal network as they could lay on several time-respecting paths due to their specific interaction dynamics.

Finally, note that although transmission centrality is not equivalent, it naturally relates to the concept of betweenness centrality (and other centrality measures based on the counts of shortest paths between nodes). As explained in details in Sect. S3 and Table S1 (Additional file 1), the difference between the two measures is rooted in their definition. While betweenness centrality considers all shortest paths between every pairs of nodes, transmission centrality takes only a single one from the potentially many other. This is especially true when $\beta = 1$ (always the case in this work), when the SI process is fully deterministic. To demonstrate these differences, we further completed a link percolation study to identify which measure, overlap or transmission centrality, is more effective to identify links connecting the network structure. Results, shown in Fig. S4(a) and discussed in the corresponding section of Additional file 1, indicates that transmission centrality provides a better strategy to identify weak ties holding the network structure together.

4.2 Case study: weak tie identification to control contagion processes in social networks

To demonstrate the potential of transmission centrality here we present a case study, where we use our new metric to identify ties in social networks in order to efficiently control contagion processes. Ties in social networks are associated with various strengths [55–57] and commonly categorized into two mutually exclusive groups: weak and strong

ties. Following the terminology introduced by Granovetter [19, 20], weak ties are maintained via sparse interactions, bridging between tightly connected communities to keep the network connected [55], and play an important role in disseminating information globally [43, 58–64]. On the other hand strong ties, sustained by frequent communications, are crucial in shaping the local connectivity of social networks, they are responsible for emerging clustered topology [62, 63, 65], and keeping information locally [43, 59, 60, 64]. A precise measure of tie strength would allow the efficient differentiation among these types and to identify weak ties in social networks in order to control globally spreading contagion processes.

Conventional measures of social tie strengths Several measures of social tie strength have lately been proposed in the literature, such as the link overlap

$$O(i, j) = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}}, \quad (2)$$

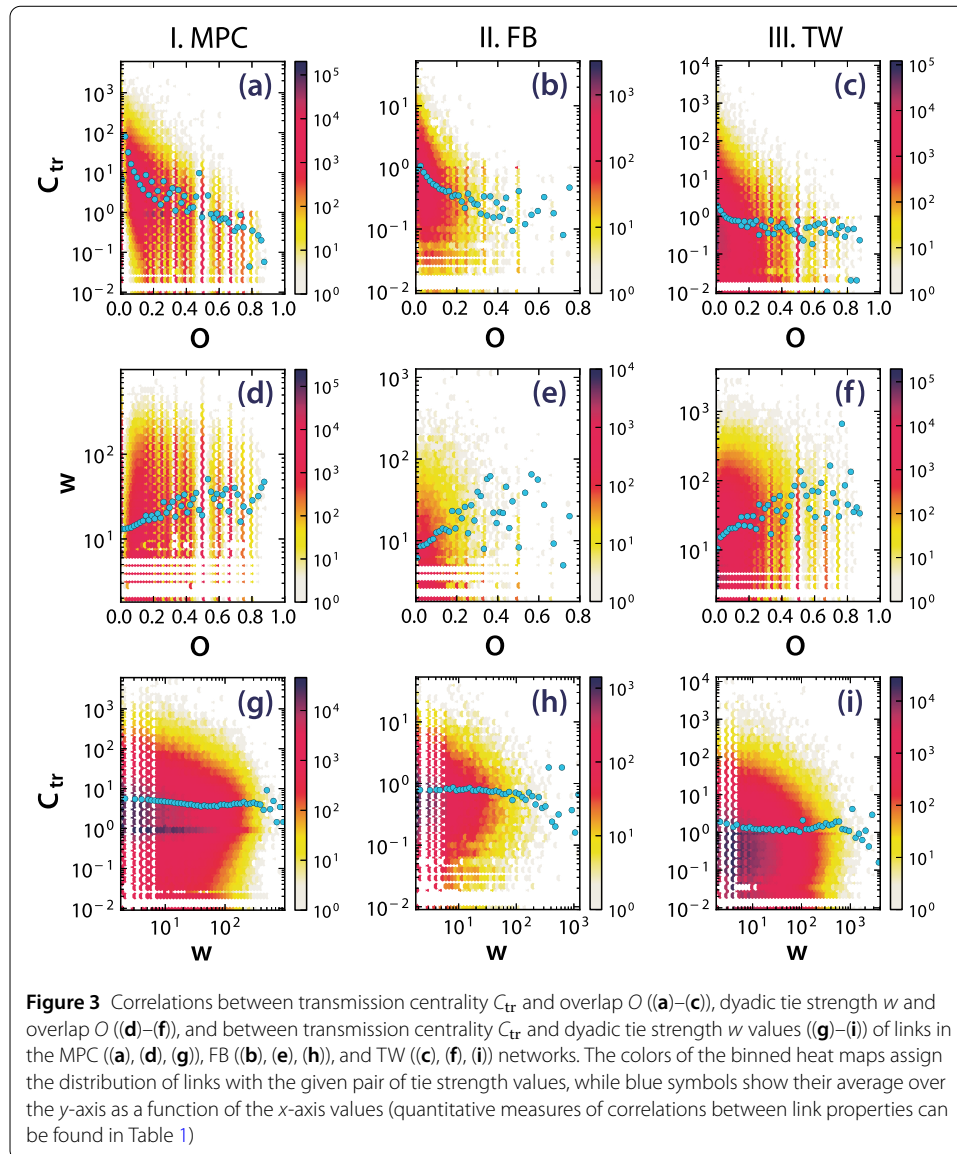
capturing the fraction of common friends in the neighborhood of connected nodes i and j [19, 55, 58]. Here, k_i and k_j assign the degree of node i and j respectively, and n_{ij} is the number of their common neighbors. Weak ties are associated with small overlap values, while the contrary is not always true. Leaf links, structural holes, or merely the fact that networks are sparse may induce links with small overlap, which leads to some ambiguity when identifying weak ties in this way.

Another way to assign the strength of social ties is via the intensity of dyadic communication [49, 58, 66]. It can be measured as the frequency, total duration, or the absolute number of interactions between connected peers. In this study, assuming discrete communication events, we define dyadic tie strength as the number of interactions between individuals i and j as

$$w(i, j) = \sum_{t=0}^T \delta(t, i, j), \quad (3)$$

where the sum runs over the observation period T . $\delta(t, i, j) = 1$ if an event appears between i and j at time t regardless of its direction, otherwise it is 0 [58]. Dyadic tie strength may capture mutual commitment or emotional closeness between people; however, as a local measure, it is subjective to individual characteristics like communication capacity or the egocentric network size. In this way, it is unable to indicate the role of a link in the global structure in the context of the emergence of any collective phenomena. In addition its broadly distributed values prohibit an evident categorization of social ties.

As shown in Fig. 3(d)–(f) and in other studies [55, 58], dyadic tie strength and link overlap are positively correlated in accordance with Granovetter’s theorem [19]. At the same time, transmission centrality and overlap show strong negative correlations (see Fig. 3(a)–(c)) as weak links, with small overlap values, are commonly situated between communities, and thus transmitting information to a large set of nodes. More interestingly, dyadic tie strength and transmission centrality values do not show strong correlations (see in Fig. 3(g)–(i)). Although both are correlated with link overlap, they capture notably different and seemingly independent features of social ties. For the precise Pearson correlation coefficients (and p -values) see Table 1.

**Table 1** Pearson correlations between transmission centrality, overlap and dyadic tie strength

Network	Pearson correlation (p -value)		
	(O, w)	(C_{tr}, O)	(C_{tr}, w)
MPC	0.097 (10^{-6})	-0.126 (10^{-6})	-0.023 (10^{-6})
FB	0.151 (10^{-6})	-0.148 (10^{-6})	-0.098 (10^{-4})
TW	0.102 (10^{-6})	-0.021 (10^{-6})	-0.002 (10^{-3})

While overlap has been shown to identify weak ties efficiently [55, 58], this measure has a major limitation. It assigns a zero overlap value for an unrealistically large fraction of links including weak ties but also leaf links, links surrounded by structural holes, or links situated at sparsely connected parts of the network. It is indeed true in the investigated systems where 48.2%, 49.8%, and 45.2% of social ties appear with $O = 0$ (resp. in the MPC, FB, TW networks). Relying merely on the link overlap one cannot differentiate between these links, thus they are treated equivalently. On the other hand, the Granovetterian crite-

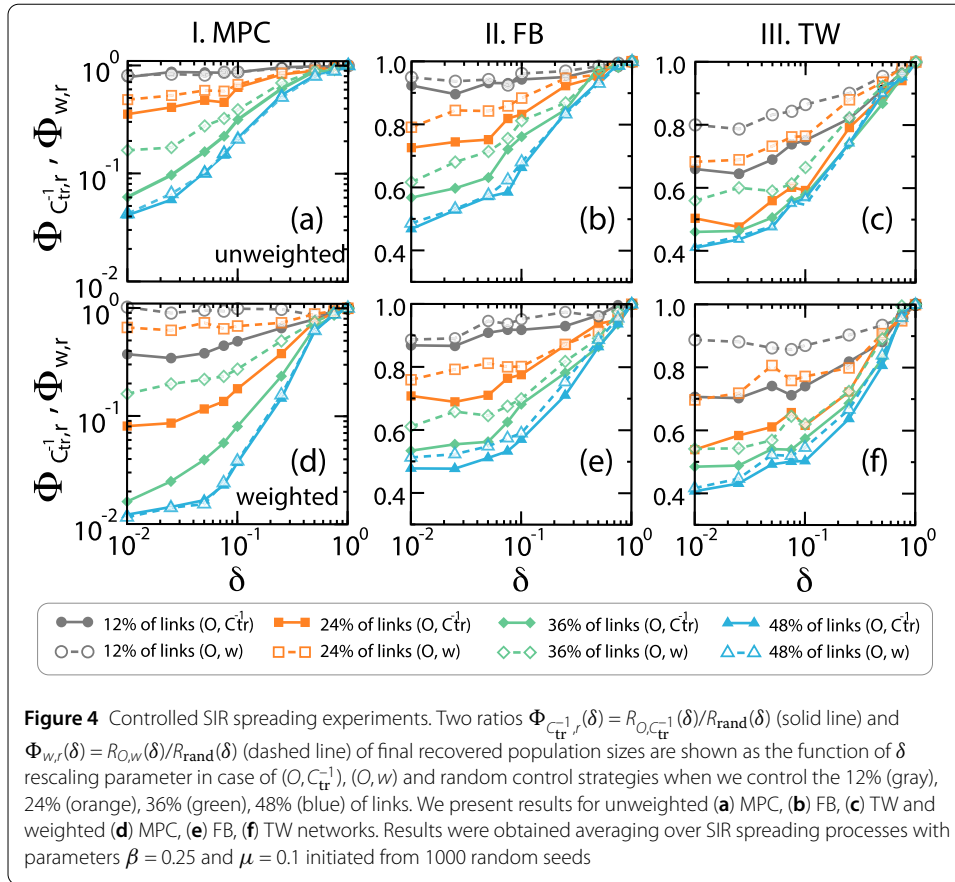
ria suggest that weak ties are not only characterized by small overlap, but they also exhibit small dyadic tie strengths, and high transmission centrality. Based on these conditions we design two combined strategies where we differentiate between zero overlap links using their w or C_{tr} values. We first rank ties in an increasing order of overlap, and then sort again links of the same overlap value increasingly by their dyadic tie strength (assigned as (O, w)), or by their inverse transmission centrality values (assigned as (O, C_{tr}^{-1})). Note, that we report a link percolation study in the Sect. S2 of Additional file 1, where we take the different single and combined weak tie measures to remove links from the network in a sorted order while measuring the average size of the remaining largest connected component. This results show that from single measures the link overlap, while from the combined measures the (O, C_{tr}^{-1}) strategy provides the best way to disconnect the network (see Fig. S2).

Controlled SIR spreading The precise identification of the weakest weak ties is important, because by suppressing interactions on this limited set of links, we may effectively control globally spreading processes in the network. To model such scenarios we take a network structure and introduce a weight ω_{ij} for each link (with values defined later). To select the weakest links to control, we consider one of the two candidate sorting strategies, (O, w) or (O, C_{tr}^{-1}) . After sorting links by one of these metrics, we select the f weakest fraction of them to control by linearly rescaling their weights as $\Omega_{ij} = \delta \omega_{ij}$, with the parameter $0 \leq \delta \leq 1$.

In this way, we weaken interactions on the selected ties, and such that we can exert further control on dynamical processes, like the Susceptible-Infected-Removed (SIR) model. The SIR process [49] is a well known model of epidemics and rumor spreading [67, 68] and it is defined on a network where nodes can be in exclusive states of susceptible (S), infected (I), or recovered (R) [49]. At each iteration connected nodes are updated as $S + I \xrightarrow{\beta} 2I$, or $I \xrightarrow{\mu} R$ with β and μ being the infection and recovery rates respectively. In this scenario, we fix $\mu = 0.1$ and $\beta = 0.25$, and re-scale the transmission probability for each controlled link as $\tilde{\beta}_{ij} = \Omega_{ij}\beta$ (for a sensitivity analysis regarding this choice see Fig. S5 (Additional file 1)). After initiating the process from a randomly selected seed we simulate it until full recovery and monitor R , the number of recovered nodes giving the number of nodes ever got infected during the process.

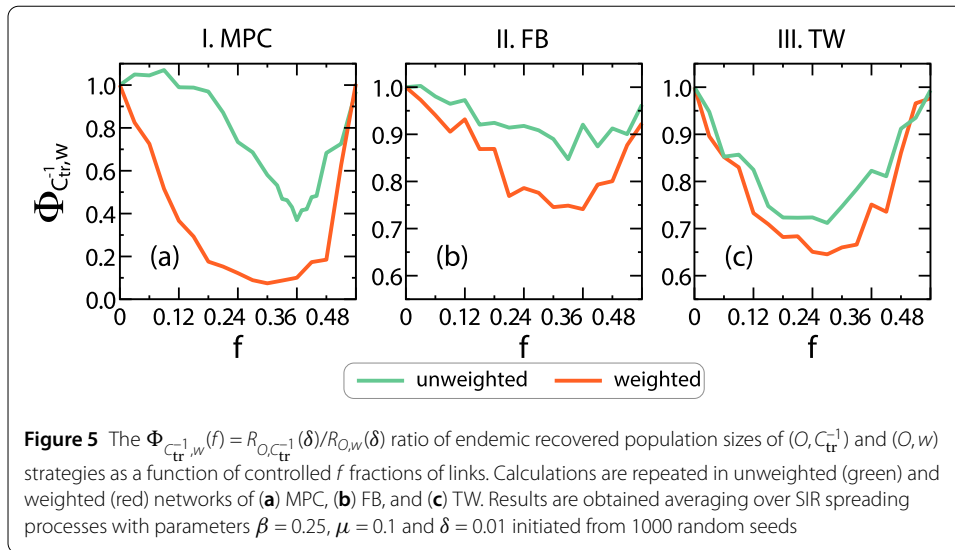
In our first experiment we assign $\omega_{i,j} = 1$ for each link assuming that the network is unweighted at the outset. To study the effects of link control, after sorting links by (O, C_{tr}^{-1}) or (O, w) , we choose the weakest 12%, 24%, 36%, or 48% of links (see Fig. 4(a), (b), and (c)). In addition, as a reference we use a network where the same fraction of randomly selected links are controlled in the same way, i.e., by re-scaling their weights with δ . Finally to quantify the effects of increasing control, we measure the $\Phi_{C_{tr}^{-1},r}(\delta) = R_{O,C_{tr}^{-1}}(\delta)/R_{rand}(\delta)$, and $\Phi_{w,r}(\delta) = R_{O,w}(\delta)/R_{rand}(\delta)$ ratios of recovered nodes in scenarios of targeted and random control strategies for various δ values. If the targeted strategy performs comparable to the random one, these ratios are equal to one; otherwise the stronger control a targeted strategy enforces, the smaller the corresponding ratio becomes. Note, that the dependency of the SIR process on the choice of its parameters is studied in Sect. S4 (Additional file 1).

When we set $\delta = 1$ the ratios of endemic population size are trivially one as no control is applied (see Fig. 4(a), (b), and (c)). However by decreasing δ , thus by increasing control, large differences appear between the targeted and random cases. Effects are stronger when



a larger fraction of weakest links are re-scaled with smaller and smaller δ factor. The differences between the (O, C_{tr}^{-1}) (solid lines) and (O, w) (dashed lines) strategies are maximal when we control an intermediate 24% or 36% of links, while they perform similarly when the controlled fraction is small (12%) or large (48%). It is also evident that the (O, C_{tr}^{-1}) strategy outperforms the (O, w) and provides remarkably better control in reducing the final infected population, specially for smaller δ values.

To bring our experiments closer to reality we repeat our measurements on weighted networks where we define link weights as $\omega_{ij} = w_{ij}/\langle w \rangle$, i.e. the number of interactions between nodes i and j normalized by the $\langle w \rangle$ average number of interactions per link calculated over the whole network. In the case where $\omega_{ij} > \langle w \rangle$ we set the corresponding weight $\omega_{ij} = 1.0$. This choice is necessary as weights are heterogeneously distributed in this case, and thus severely slow down the simulated spreading to reach full prevalence. On the other hand, since controlled links with small overlap values tend to have small weights, negligible effect of this approximation is expected. The different control strategies qualitatively provide the same results on the weighted FB and TW networks (Fig. 4(e), (f)); however, their effects are considerably stronger on the MPC structure (Fig. 4(d)). There, the (O, C_{tr}^{-1}) strategy appears to be the more efficient even after controlling only the 12% of the ties. Moreover, this strategy can lead to 90% reduction of the infected population in the case of re-scaling 36% of links with $\delta = 0.01$. Note that the observed differences between different strategies cannot be the result of the limited communication on zero overlap links only, as we observed qualitatively the same effects in weighted and unweighted networks.



To directly highlight the differences between the targeted strategies we further investigate the strongest controlled case. We set $\delta = 0.01$ and repeat our experiments by controlling various f fractions of links to measure the $\Phi_{C_{tr}^{-1},w}^{-1}(f) = R_{O,C_{tr}^{-1}}(\delta)/R_{O,w}(\delta)$ fraction of endemic recovered population sizes, i.e., the ratio of the two performance functions. Results in Fig. 5(a), (b), and (c) evidently show that the (O, C_{tr}^{-1}) strategy almost always outperforms the (O, w) strategy, especially when we consider weights. In addition, the minimum points of the $\Phi_{C_{tr}^{-1},w}^{-1}(f)$ curves in Fig. 5 assign the best pay-off between the controlled f fraction of links and the effectiveness of contamination control using the (O, C_{tr}^{-1}) strategy. This minimum point indicates that $\sim 30\%$ of the weakest ties are enough to control and mostly efficiently hinder the spreading processes on the investigated social networks. Note, that we performed similar experiments to measure $\Phi_{C_{tr}^{-1},C_b}^{-1}(f) = R_{O,C_{tr}^{-1}}(\delta)/R_{O,C_b}(\delta)$ ratio, which compares the performance of strategies using combined measures of overlap, transmission centrality and betweenness centrality. Results on the Facebook network shown in Fig. S4(b) indicates that transmission centrality outperforms betweenness centrality in this matter as well.

5 Discussion

In this study we introduced a new link centrality measure, called *transmission centrality*, which sensitively quantifies the importance of links in global diffusion processes. We defined an algorithm to compute transmission centrality, demonstrated on three large-scale networks its general properties, and discussed possible ways of how this measure can be generalized for directed, weighted or temporal networks or even as a node centrality measure. Finally in a case study, we showed that the combined information of overlap and transmission centrality serves as the best way to identify weak links to gain maximum control of spreading processes. Although here we demonstrated the effectiveness of transmission centrality in identifying weak ties in social networks specifically, the same metric can be applied in any other type of networks to identify links with specific structural role and importance in controlling the emergence of various collective phenomena.

We discussed that the main limitation of this new centrality measure is rooted in its computational complexity, which scales as the best known algorithm for betweenness centrality. However, we proposed a way around this limitation by defining a heuristic method

to approximate transmission centrality values in very large networks at a considerably cheaper cost.

Most of earlier methods to control spreading processes were focusing on influential nodes as their removal provided efficient ways to hinder epidemics. Controlling links in a network is in a way more expensive process but on the other hand it provides the advantage to control epidemics without isolating nodes (e.g. a person) from the rest of the network but only from a limited number of neighbors. At the same time, the control of a large fraction of links in a social network is virtually impossible. This is where our method provides an advantage by indicating the minimum set of most important links to control in order to suppress epidemics effectively.

Several extensions of this method are possible by considering other probing processes other than SI process, or arbitrary weight definitions, directed links, temporal interactions, or node transmission centrality. Furthermore, several straightforward applications can be foreseen. Examples are in viral marketing, rumor contamination, or intervention designs; their identification can be the subject of future studies. Our aim here is to ground a new metric of link centrality and to contribute to the design of effective methods to identify ties, which play an indisputably important role in the structure and dynamics of social networks.

Additional material

Additional file 1: The supplementary information includes additional results on radius bias analysis, correlations between link transmission centralities and link betweenness centralities, and sensitivity analysis for controlling weak ties experiment. (PDF 1.6 MB)

Acknowledgements

We are grateful for D. Mocanu for her help in data preparation and N. Samay for her help in visualization. QZ would like to acknowledge Dr. Duygu Balcan for mentorship and invaluable contributions in the beginning of this project.

Funding

This project has been supported by the SoSweet ANR project (ANR-15-CE38-0011-03).

List of abbreviations

MPC, mobile phone communication network; FB, Facebook wall post interaction network; TW, twitter conversation network; SI, susceptible-infected model; SIR, susceptible-infected-recover model; LCC, the largest connected component.

Availability of data and materials

The data of wall posts of Facebook users are publicly accessible from [45–47]. The Twitter conversation network data could be available upon request. The mobile phone call data was shared after the signature of several non-disclosure agreements between the authors and the provider. Even the dataset is anonymized, it may contain personally sensitive informations, which cannot be shared publicly to secure to privacy of the users. Access to the data for verification purposes may be granted upon request and only within the secured facilities of the hosting institute.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MK, AV and QZ designed the study. QZ and MK performed experiments and data analysis. All authors contributed to writing the manuscript. All authors read and approved the final manuscript.

Author details

¹Laboratory for the Modeling Biological and Socio-technical Systems, Northeastern University, Boston, USA. ²LIP UMR 5668, IXXI, Université de Lyon, ENS de Lyon, Inria, CNRS, UCB Lyon 1, Lyon, France.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 14 February 2018 Accepted: 31 August 2018 Published online: 14 September 2018

References

- Freeman LC (1977) A set of measures of centrality based on betweenness. *Sociometry* 40(1):35–41
- Bavelas A (1950) Communication patterns in task-oriented groups. *J Acoust Soc Am* 22(6):725–730
- Katz L (1953) A new status index derived from sociometric analysis. *Psychometrika* 18(1):39–43
- Bollobás B, Erdős P (1984) Graph theory and combinatorics: proceedings of the Cambridge combinatorial conference in honour of Paul Erdős. Academic Press, Cambridge
- Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. *Nat Phys* 6(11):888–893
- Estrada E, Rodriguez-Velazquez JA (2005) Subgraph centrality in complex networks. *Phys Rev E* 71(5):056103
- Everett MG, Borgatti SP (2010) Induced, endogenous and exogenous centrality. *Soc Netw* 32(4):339–344
- Brandes U (2001) A faster algorithm for betweenness centrality. *J Math Sociol* 25(2):163–177
- Ercsey-Ravasz M, Toroczkai Z (2010) Centrality scaling in large networks. *Phys Rev Lett* 105:038701
- Borgatti SP (2005) Centrality and network flow. *Soc Netw* 27(1):55–71
- Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. *Comput Netw ISDN Syst* 30(1–7):107–117
- Leontief WW (1941) The structure of American economy, 1919–1929: an empirical application of equilibrium analysis. Harvard University Press, Cambridge
- Travençolo BAN, Costa LdF (2008) Accessibility in complex networks. *Phys Lett A* 373(1):89–95
- Lawyer G (2015) Understanding the influence of all nodes in a network. *Sci Rep* 5:8665
- Piraveenan M, Prokopenko M, Hossain L (2013) Percolation centrality: quantifying graph-theoretic impact of nodes during percolation in networks. *PLoS ONE* 8(1):53095
- Bajardi P, Poletto C, Ramasco JJ, Tizzoni M, Colizza V, Vespignani A (2011) Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. *PLoS ONE* 6(1):16591
- Christakis NA, Fowler JH (2010) Social network sensors for early detection of contagious outbreaks. *PLoS ONE* 5(9):12948
- Gemmetto V, Barrat A, Cattuto C (2014) Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC Infect Dis* 14(1):695
- Granovetter MS (1973) The strength of weak ties. *Am J Sociol* 78:1360–1380
- Granovetter MS (1983) The strength of weak ties: a network theory revisited. *Sociol Theory* 1:201–233
- Newman M (2010) Networks: an introduction. Oxford University Press, Oxford
- Brandes U, Pich C (2007) Centrality estimation in large networks. *Int J Bifurc Chaos* 17(7):2303–2318
- Brandes U (2008) On variants of shortest-path betweenness centrality and their generic computation. *Soc Netw* 30(2):136–145
- Bader DA, Kintali S, Madduri K, Mihail M (2007) Approximating betweenness centrality. In: WAW, vol 4863. Springer, Berlin, pp 124–137
- Geisberger R, Sanders P, Schultes D (2008) Better approximation of betweenness centrality. In: Proceedings of the meeting on algorithm engineering & experiments. Society for Industrial and Applied Mathematics, Philadelphia, pp 90–100
- Riondato M, Kornaropoulos EM (2016) Fast approximation of betweenness centrality through sampling. *Data Min Knowl Discov* 30(2):438–475
- Jensen P, Morini M, Karsai M, Venturini T, Vespignani A, Jacomy M, Cointet J-P, Mercklé P, Fleury E (2016) Detecting global bridges in networks. *J Complex Netw* 4(3):319–329
- Alahakoon T, Tripathi R, Kourtellis N, Simha R, Iamnitchi A (2011) K-path centrality: a new centrality measure in social networks. In: Proceedings of the 4th workshop on social network systems. ACM, New York
- Chen W, Teng S-H (2017) Interplay between social influence and network centrality: a comparative study on shapley centrality and single-node-influence centrality. In: Proceedings of the 26th international conference on world wide web, pp 967–976
- Rossi MEG, Vazirgiannis M (2016) Exploring network centralities in spreading processes. In: International symposium on web algorithms (iSWAG)
- Liu J-G, Lin J-H, Guo Q, Zhou T (2016) Locating influential nodes via dynamics-sensitive centrality. *Sci Rep* 6:21380
- Teixeira AS, Monteiro PT, Carriço JA, Ramirez M, Francisco AP (2013) Spanning edge betweenness. In: Workshop on mining and learning with graphs, vol 24, pp 27–31
- Mavroforakis C, Garcia-Lebron R, Koutis I, Terzi E (2015) Spanning edge centrality: large-scale computation and applications. In: Proceedings of the 24th international conference on world wide web, pp 732–742
- De Meo P, Ferrara E, Fiumara G, Ricciardello A (2012) A novel measure of edge centrality in social networks. *Knowl-Based Syst* 30:136–150
- De Meo P, Ferrara E, Fiumara G, Provetti A (2014) On Facebook, most ties are weak. *Commun ACM* 57(11):78–84
- Everett MG, Valente TW (2016) Bridging, brokerage and betweenness. *Soc Netw* 44:202–208
- Lü L, Chen D, Ren X-L, Zhang Q-M, Zhang Y-C, Zhou T (2016) Vital nodes identification in complex networks. *Phys Rep* 650:1–63
- Gu J, Lee S, Saramäki J, Holme P (2017) Ranking influential spreaders is an ill-defined problem. *Europhys Lett* 118(6):68002
- Cheng X-Q, Ren F-X, Shen H-W, Zhang Z-K, Zhou T (2010) Bridgeness: a local index on edge significance in maintaining global connectivity. *J Stat Mech Theory Exp* 2010(10):10011
- Cui A-X, Yang Z, Zhou T (2016) Strong ties promote the epidemic prevalence in susceptible–infected–susceptible spreading dynamics. *Phys A, Stat Mech Appl* 445:335–342
- Seidman SB (1983) Network structure and minimum degree. *Soc Netw* 5(3):269–287
- Bollobás B, Erdős P (1984) Graph theory and combinatorics: proceedings of the Cambridge combinatorial conference in honour of Paul Erdős. Academic Press, Cambridge
- Karsai M, Kivela M, Pan RK, Kaski K, Kertész J, Barabási A-L, Saramäki J (2011) Small but slow world: how network topology and burstiness slow down spreading. *Phys Rev E* 83:025102
- Kivela M, Pan RK, Kaski K, Kertész J, Saramäki J, Karsai M (2012) Multiscale analysis of spreading in a large communication network. *J Stat Mech Theory Exp* 2012(3):03005

45. Facebook wall posts network dataset—KONECT. <http://konect.uni-koblenz.de/networks/facebook-wosn-wall> (2014)
46. Viswanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user interaction in Facebook. In: Proceedings of the 2nd ACM workshop on online social networks. ACM, New York, pp 37–42
47. Kunegis J (2013) KONECT: the Koblenz network collection. In: Proceedings of the international web observatory workshop, pp 1343–1350
48. Guide to the Twitter API part 3 of 3: an overview of Twitters streaming API. <http://blog.gnip.com/tag/gardenhose/> (2014)
49. Barrat A, Barthélemy M, Vespignani A (2008) Dynamical processes on complex networks. Cambridge University Press, Cambridge
50. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms, 2nd edn. MIT Press, Cambridge
51. Rodriguez-Iturbe I, Rinaldo A (2001) Fractal river basins: chance and self-organization. Cambridge University Press, Cambridge
52. Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. *SIAM Rev* 51(4):661–703
53. Holme P, Saramäki J (2012) Temporal networks. *Phys Rep* 519(3):97–125
54. Karsai M, Perra N, Vespignani A (2014) Time varying networks and the weakness of strong ties. *Sci Rep* 4:4001
55. Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási A-L (2007) Structure and tie strengths in mobile communication networks. *Proc Natl Acad Sci USA* 104:7332–7336
56. Saramäki J, Leicht E, López E, Roberts SG, Reed-Tsochas F, Dunbar RI (2014) Persistence of social signatures in human communication. *Proc Natl Acad Sci USA* 111(3):942–947
57. Palchykov V, Kaski K, Kertész J, Barabási A-L, Dunbar RI (2012) Sex differences in intimate relationships. *Sci Rep* 2:370
58. Onnela J-P, Saramäki J, Hyvönen J, Szabó G, de Menezes MA, Kaski K, Kertész J, Barabási A-L, Kertész J (2007) Analysis of a large-scale weighted network of one-to-one human communication. *New J Phys* 9:179
59. Centola D, Macy M (2007) Complex contagions and the weakness of long ties. *Am J Sociol* 113(3):702–734
60. Centola D (2010) The spread of behavior in an online social network experiment. *Science* 329(5996):1194–1197
61. Ghasemiesfeh G, Ebrahimi R, Gao J (2013) Complex contagion and the weakness of long ties in social networks: revisited. In: Proceedings of the fourteenth ACM conference on electronic commerce. EC '13, pp 507–524
62. Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. *Science* 311(5757):88–90
63. Kumpula JM, Onnela J-P, Saramäki J, Kaski K, Kertész J (2007) Emergence of communities in weighted networks. *Phys Rev Lett* 99(22):228701
64. Miritello G, Moro E, Lara R (2011) Dynamical strength of social ties in information spreading. *Phys Rev E* 83(4):045102
65. Rapoport A (1953) Spread of information through a population with socio-structural bias: I. Assumption of transitivity. *Bull Math Biophys* 15(4):523–533
66. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. *Proc Natl Acad Sci USA* 101(11):3747–3752
67. Anderson RM, May RM (1992) Infectious diseases of humans: dynamics and control. Oxford University Press, Oxford
68. Nekovee M, Moreno Y, Bianconi G, Marsili M (2007) Theory of rumour spreading in complex social networks. *Phys A, Stat Mech Appl* 374:457–470

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com
