Supplementary Information:
Link transmission centrality in large-scale social networks

Qian Zhang, Márton Karsai, Alessandro Vespignani

S1 Local bias of link transmission centrality

In the main text, we have discussed that link transmission centrality is a locally biased measure as it assigns higher values to links, which are closer to the seed node. To understand this bias on the "close-to-seed" links, we directly analyze the actual branching trees with root as the actual seed node. We compute the transmission centrality C_{tr} of links as the function of their distance d from the actual root. Here the distance of a link is defined as $d = min(\ell_b, \ell_e)$, i. e., the minimum of the shortest paths ℓ_b or ℓ_e of the beginning and ending nodes (respectively) of the actual link.

One can assume two different characters for links in the vicinity of the seed node. First, if the seed has low degree, the corresponding C_{tr} values of immediate links will be larger compared to the case where the seed has many neighbors. Second, we expect that this bias decreases by distance d measured from the seed. These two effects can be identified from the scaling of the $C_{tr}(d)$ functions in Fig.S1.a-c. Here for each network we select randomly seeds from different degree groups (100 seeds for each group) and measure the C_{tr} average transmission centrality (Fig.S1.a-c) in distance smaller than or equal to d relative to the actual seed.

These results can help us estimate the induced bias and select an appropriate threshold d_{max} to eliminate its effect. This choice has to consider two competing factors: the choice of a d_{max}, which is large enough that the local bias of the seed becomes negligible; and to choose a distance small enough not to remove too many links from the tree. The E number of links in distance d from the seed is exponentially increasing as shown in (Fig.S1.d-f). To remove the actual bias we set $C_{tr} = 0$ for those links, which are within a determined distance d_{max} from the actual seed in the network. Based on Fig.S1 we choose $d_{max} = 8$, $d_{max} = 3$, and $d_{max} = 7$, for the MPC, FB and TW datasets respectively. These choices fulfill both conditions as they are large enough to reduce the bias considerably, while at the same time exclude only the 4.3%, 1.1% and 3.7% of links for MPC, FB and TW datasets respectively. Naturally, removing links may decrease the overall heterogeneity of C_{tr}. However, as demonstrated in Fig.S1.g-i, even after excluding the biased links, the distribution $P(C_{tr})$ remains fat tailed.

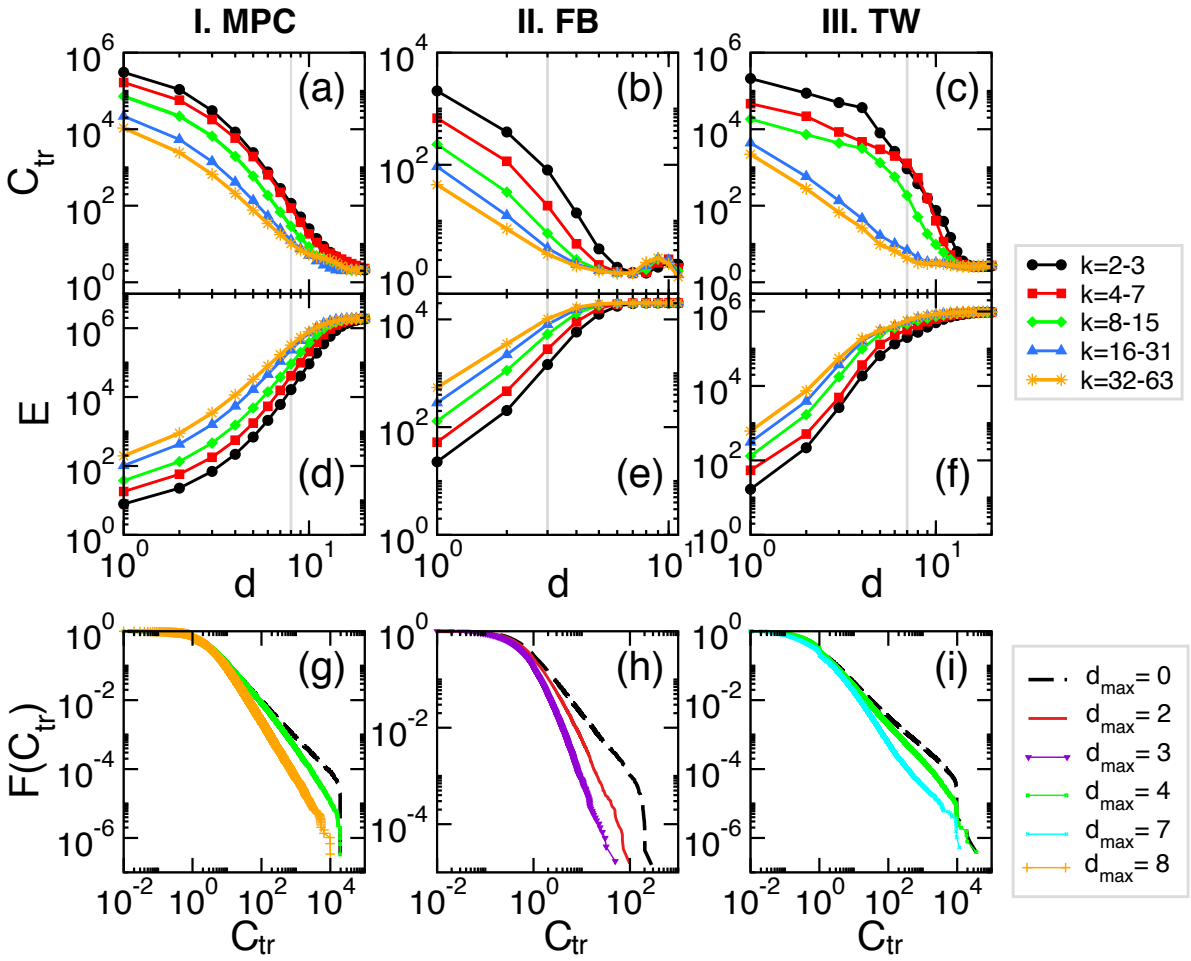

Figure S1: Local bias of link transmission centrality on the MPC, the FB and the TW networks. (a,b,c) Average importance C_{tr} of links in distance d from the actual seed for the three network respectively. (d,e,f) Average number of links in distance d from the actual seed. Random seeds were selected from different degree groups (100 seeds for each group). (g,h,i) Distribution of transmission centrality values after unbiasing with links in distance d.

S2 Percolation analysis

As weak ties [2, 3] are commonly situated between densely connected parts of the social network, their removal lead to a rapid segmentation of the structure. We use this condition to identify the best tie strength measure in indicating weak ties, which are the most effective in decoupling the network structure. To do so, (a) we calculate tie strength for each link (b) sort them in an increasing order, (c) remove an f fraction of them, and (c) monitor the remaining structure. We measure quantities borrowed from percolation theory

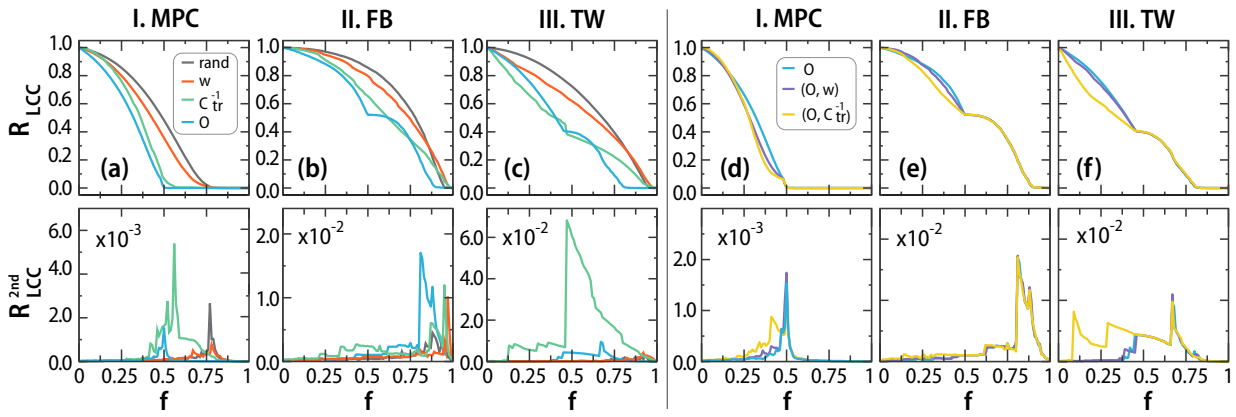

Figure S2: Link removal analysis of the (a,d) MPC, (b,e) FB, and TW (c,f) networks. Upper (lower) panels depict the R_{LCC} (R^{2nd}_{LCC}) fraction of nodes in the largest (2nd largest) connected components after an f fraction of links were removed. In panels (a,b,c) links were removed in a sorted order ranked randomly (gray line), by their dyadic tie strength w (orange line), by their inverse transmission centrality C^{-1}_{tr} (green line), or by link overlap O (blue line). In panels (d,e,f) links were removed in a combined sorted order ranked first by their O and then by their dyadic tie strength (O, w) (purple line) or by their inverse transmission centrality (O, C^{-1}_{tr}) values (yellow line)

[5] like $R_{LCC}(f) = N_{LCC}(f)/N$ relative size of the LCC, which indicates the global connectedness of the structure and goes to zero as the network becomes fragmented. But we also measure $R^{2nd}_{LCC}(f)$, the relative size of the second largest connected component, which has a maximum (divergent in the thermodynamic limit) when the network falls apart. The better the actual tie strength measure identifies weak links, the faster the fragmentation appears during the link removal process [4].

At first, to sort links we follow four ranking strategies: we remove links sorted increasingly by their dyadic tie strength w; by the inverse of their transmission centrality values C^{-1}_{tr}; by their overlap values O; or as a reference we remove links in a random order. As seen in Fig.S2.a-c (upper panels), in all investigated systems the random removal strategy (gray line) is the least effective to segment the network, and even dyadic tie strength w (orange line) proposes a faster decomposition, the segmentation points of these two strategies were matching closely (see lower panels). This demonstrates that dyadic tie strength as a local measure cannot capture effectively links, which are responsible for the global connectedness of the structure. A different scenario appears if we remove links ranked by their transmission centrality or overlap values. Here the decomposition evolves much faster and the structure falls apart earlier suggesting that C_{tr} and O provide more efficient ways to identify links bridging between communities. Note that the elbows in the O based percolation curves in Fig.S2.b and c (upper panels) are corresponding to the typically size of communities, which are commonly larger in online social networks.

Even overlap seems to be the most effective metric to identify weak ties, this measure has a major limitation. It assigns a zero overlap vale to an unrealistically large fraction of links, thus providing no further way of differentiation between them. Consequently these ties are treated equivalently and removed in a random order in this case. It is indeed true in the investigated systems where the 48.2%, 49.8%, and 45.2% of social ties appear with $O = 0$ in the MPC, FB, TW networks respectively. On the other hand,

the Granovetterian criteria suggest that a weak tie is not only characterised by a small overlap, but it also exhibits small dyadic tie strength, and high transmission centrality. Based on these conditions we can design two combined strategies where we differentiate between zero overlap links using their other w or C_{tr} values. Here, we first rank ties in an increasing order of overlap, and then remove links of the same overlap value increasingly by their dyadic tie strength (assigned as (O, w)) or inverse transmission centrality values ((O, C_{tr}^{-1})). Following these combined link removal strategies we found (in Fig.S2.d, e, and f) that, even the improvement is minor in some cases, yet the (O, C_{tr}^{-1}) strategy propose the most efficient way to decouple the global structure and to decrease the relative size of the LCC. Consequently transmission centrality provides the best way to label zero overlap links to identify the weakest weak ties in the network structure.

S3 Relation to betweenness centrality

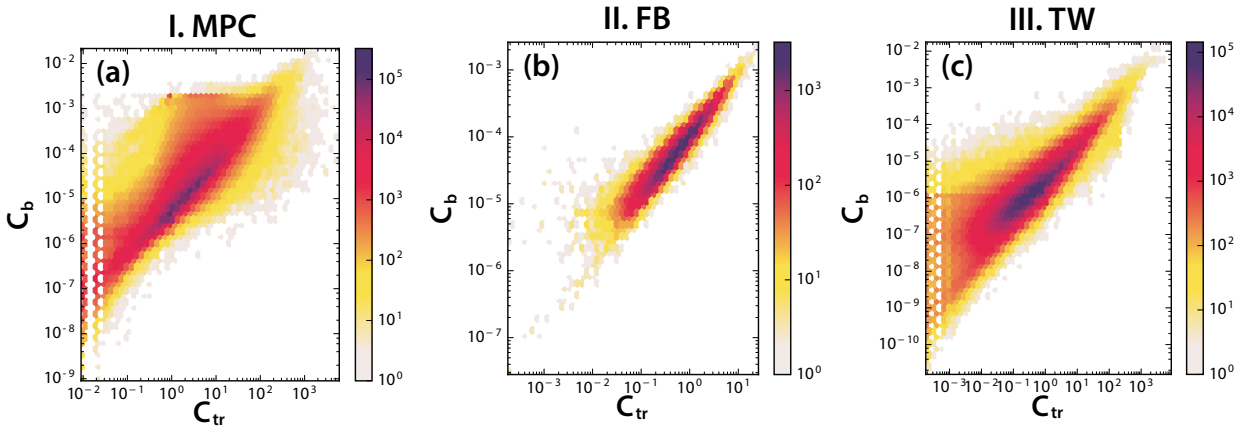

Figure S3: (a,b,c) Correlations between sampled and unbiased link transmission centrality C_{tr} and link betweenness centrality C_b values in case of the MPC (Pearson $r = 0.83$, p-value $< 10^{-6}$), FB (Pearson $r = 0.97$, p-value $< 10^{-6}$), and TW (Pearson $r = 0.91$, p-value $< 10^{-6}$) networks respectively. Computations of C_{tr} were initiated from 2000 (5000 and 5000) seeds in each case (respectively). For the MPC network C_b values were calculated between 5000 randomly selected nodes and any other nodes in the network.

Transmission centrality as a measure can be easily associated to the concept of link betweenness centrality commonly defined as $C_b(i,j) = \sum_{s\neq i\neq j\neq t} g_{s,t}(i,j)/g_{s,t}$, where $g_{s,t}$ assigns the number of shortest paths between nodes s and t while $g_{s,t}(i,j)$ is the number of them, which goes through the link (i,j). Although the definition of C_b and C_{tr} are not equivalent they are strongly related. Their differences are rooted in the deterministic definition of C_b, which considers all shortest paths between pairs of nodes. On the other hand, C_{tr} is defined by an SI process, which is stochastic even in case of $\beta = 1$, as it considers in a random order the possible links of an infected node along which to transmit the infection to susceptible neighbours. In this way it may never explore *all possible shortest paths* but would give credit to the most plausible ones over several attempts of realizations. Despite these fundamental differences they both capture similar

quantities fractional to the number of shortest paths running through a given link. Due to these underlying similarities they appear to be closely correlated in the case of all investigated empirical structures as seen in Fig.S3.

Table S1: The r Pearson correlation coefficients measured between the betweenness centrality and transmission centrality values of the same links with increasing number of seeds in the three empirical network. Each corresponding p-value is smaller than 0.001.

No. Seed	10	50	100	500	1000	2000
r_{MPC}	0.16	0.27	0.48	0.67	0.79	0.83
r_{FB}	0.36	0.58	0.74	0.92	0.94	0.96
r_{TW}	0.40	0.55	0.65	0.76	0.84	0.87

These strong correlations demonstrate the relationship between transmission centrality and betweenness centrality, with the advantage that the approximate calculation of the former scales considerably better with system size. As discussed earlier, the exact computation of transmission centrality scales with $\mathcal{O}(|V||E|)$ complexity, which is equal to the best known algorithm [1] to measure C_b. On the other hand, one can reduce radically the computational cost by considering a relatively small number of seeds to compute average C_{tr} and to obtain surprisingly good approximations for link betweenness values. To demonstrate this scaling, for all three empirical networks we measured the correlations between C_b and C_{tr} while successively increasing the number of seeds for the latter one. Results summarized in Table S1 show that the average transmission centrality, computed from the 0.1% of the MPC network nodes (10% in case of FB and 0.2% for TW), approximates well the actual betweenness centrality values, with correlations $R_{MPC} \simeq 0.83$ (resp. $R_{FB} \simeq 0.96$ and $R_{TW} \simeq 0.87$). This demonstrates yet again the close relationship between these metrics and the success of the provided approximation method of transmission centrality.

To further compare the effectiveness of betweenness centrality and transmission centrality in terms of maintaining connectivity and controlling epidemic spreading process, we calculated link betweenness centrality for all links on Facebook wall post network. In Fig.S4 (a), we show percolation analysis following the same procedure in Sec.S2, and monitor the global connectedness of the network structure by removing links ordered by transmission centrality (dotted blue line), betweenness centrality (dotted yellow line), combination of overlap and transmission centrality (blue line), and combination of overlap and betweenness centrality (yellow line). It it clear that the decomposition evolves much faster when the links are removed following ordered transmission centralities than following ordered betwenness centralities, although the structure falls apart at the same fraction of removed links. Similar patterns are observed when the links are removed following the combination strategies. Thus, from perspective of percolation analysis, betwenness centrality does not perform better than transmission centrality.

In Fig.S4 (b), we also quantify the effectiveness of controlling SIR processes based on two strategies of identifying weak ties. The presented simulation results used a single parameter set with a constant basic reproduction number of $R_0 = \beta/\mu = 2.5$. We monitor $\Phi_{C_{tr}^{-1},C_b}(f) = R_{O,C_{tr}^{-1}}(f)/R_{C_b}(f)$ ratio between the endemic recovered population sizes after scaling the weights of links selected by the (O, C_{tr}^{-1}) strategy and the (O, C_b) strategy. $\Phi_{C_{tr}^{-1},C_b} < 1$ indicates that controlling weak ties identified by C_{tr}^{-1} is more effective than controlling weak ties identified by C_b. As shown in the figure, the strategy of controlling weak ties identified by C_b does not perform more effective than the strategy of C_{tr}^{-1} from the perspective of the final

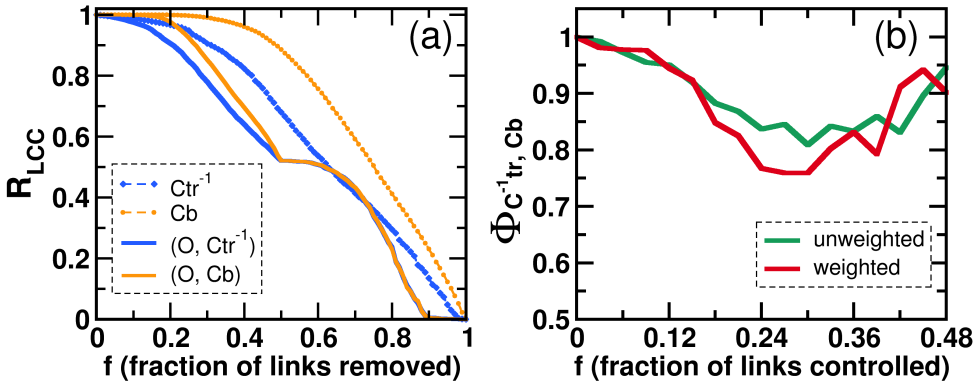

Figure S4: Comparison of the performance of transmission centrality and betweenness for Facebook wall post network. (a) percolation analysis of monitoring the relative size of the largest connected component as a function of removed links. (b) controlling SIR spreading process on weak ties selected by two centrality metrics (O, C_{tr}^{-1}) and (O, C_b).

epidemic sizes. Therefore, regardless of computational costs, the transmission centrality performs better than betweenness centrality in terms of maintaining connectedness of the network structure and effectively reducing epidemic sizes.

S4 Parameter dependence of controlled SIR spreading

In the main text we argued that the combined metrics of overlap-transmission centrality (O, C_{tr}^{-1}) is the most efficient metric to hinder epidemics outbreaks modeled by an SIR process. However, all presented simulation results used a single parameter set with a constant basic reproduction number of $R_0 = \beta/\mu = 2.5$. Here, to demonstrate that our simulation results were mostly independent of the choice of R_0, we fixed $\mu = 0.1$ and repeated our experiments for different values of β. We selected an f fraction of links by the (O, C_{tr}^{-1}) strategy or randomly, and re-scaled their weight by $\delta = 0.01$. We measured the average final size of the recovered population through 100 realizations for the targeted and random strategies. Depicting the ratio $\Phi_{C_{tr}^{-1},r}(f) = R_{O,C_{tr}^{-1}}(f)/R_{random}(f)$ of the corresponding measures, in Fig.S5, we show that the effects of targeted control increases as we control a higher fraction of links (just as we have seen in the main text); however this behaviour depends only weakly on the choice of β. This suggests that the observed behaviour is qualitatively the same for a wide range of R_0 of the SIR model, thus not a consequence of specific parametrization of the spreading process.

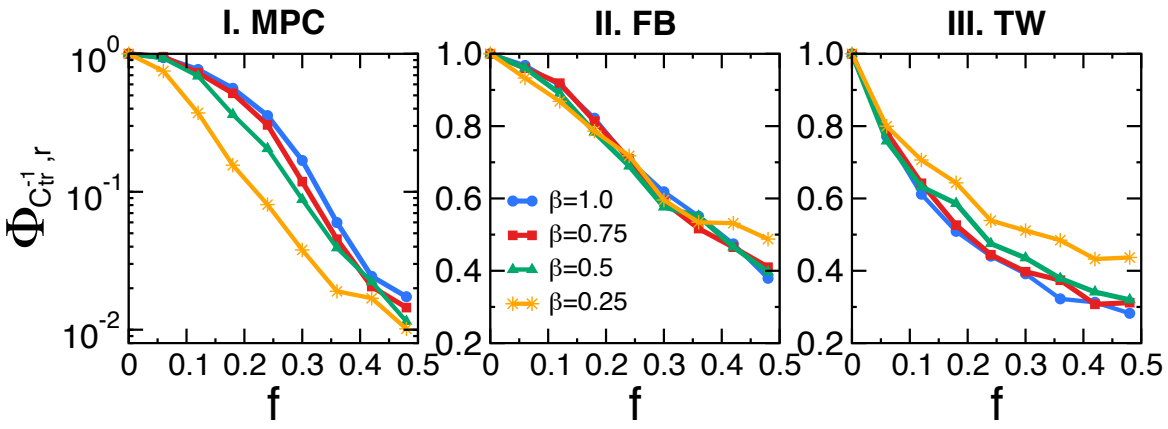

Figure S5: Parameter dependence of SIR spreading. We control weak ties to impede SIR spreading on (a) the MPC, (b) the FP, and (c) the TW networks with fixed scaling factor $\delta = 0.01$. The recovery rate μ of the SIR process is fixed to 0.1. We show the $\Phi_{C_{tr}^{-1},r}(f) = R_{O,C_{tr}^{-1}}(f)/R_{rand}(f)$ ratio between the endemic recovered population sizes after scaling the weights of links selected by the (O, C_{tr}^{-1}) strategy and randomly. We simulate the process for different values of the infection rate β (1.0 blue dots, 0.75 red squares, 0.5 green triangles, and 0.25 orange stars) as the function of the f fraction of controlled links.

References

[1] Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008.

[2] Mark S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.

[3] Mark S. Granovetter. The strength of weak ties: A network theory revisited. *Sociological theory*, 1:201–233, 1983.

[4] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. de Menezes, K. Kaski, J. Kertész, A.-L. Barabási, and J Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9:179, 2007.

[5] Dietrich Stauffer and Amnon Aharony. *Introduction to Percolation Theory*. Taylor & Francis, 2003.