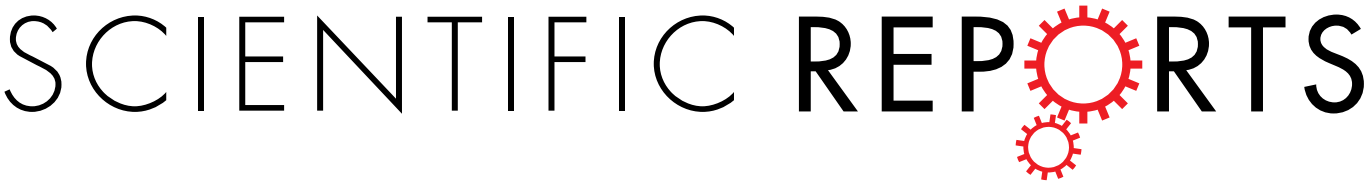

OPEN

Super-Spreader Identification Using Meta-Centrality

Andrea Madotto & Jiming Liu

Received: 12 April 2016
Accepted: 15 November 2016
Published: 23 December 2016

Super-spreaders are the nodes of a network that can maximize their impacts on other nodes, e.g., in the case of information spreading or virus propagation. Many centrality measures have been proposed to identify such nodes from a given network. However, it has been observed that the identification accuracy based on those measures is not always satisfactory among different types of networks. In addition, the nodes identified by using single centrality are not always placed in the top section, where the super-spreaders are supposed to be, of the ranking generated by simulation. In this paper we take a meta-centrality approach by combining different centrality measures using a modified version of Borda count aggregation method. As a result, we are able to improve the performance of super-spreader identification for a broad range of real-world networks. While doing so, we discover a pattern in the centrality measures involved in the aggregation with respect to the topological structures of the networks used in the experiments. Further, we study the eigenvalues of the Laplacian matrix, also known as Laplacian spectrum, and by using the Earth Mover's distance as a metric for the spectrum, we are able to identify four clusters to explain the aggregation results.

The super-spreaders are the nodes in a network that can maximize their impacts on other nodes, as in the case of information spreading or virus propagation. The identification of these nodes is very important in many real-world domains, ranging from innovation diffusion[1], viral marketing[2], to epidemic disease identification and control[3–5]. Many methods, in particular centrality measure based methods, have been proposed in the past that aim to model and identify the most influential spreaders in complex networks. Among them, K-shell Decomposition method[6–8] and Expected Force method[9] have shown better performance than others under various epidemiological models. To evaluate the accuracy of the aforementioned measures various techniques has been employed. The most common one makes use of epidemic simulations[10] from a single seed node, while the average infection size acts as the spreading characterization of the seed. Another possible way of evaluation is based on real-life tracking (e.g., in information spreading[11]) but, unfortunately, this approach is not always applicable due to the lack of data, for example, in transportation and computer networks. On the other hand, there are many networks where it is possible to quantitatively estimate the probability of spreading at the node level (e.g., based on such information as the number of passengers in a flight or number co-authorship papers) and bind it to the strength of connection between the nodes as a weight value associated with the edge[10]. This way allows for modeling and simulating the impacts of a spreader by taking into consideration the probability of node-level spreading. For instance, in the case of modeling information diffusion in a social network, the individual or meta-population level contact information (e.g., the duration or the frequency of the contact) can play an important role; this information will be reflected in the edge weights of the network[12].

In essence, a centrality measure computes ranking score for the nodes of a network based on their connectivity characteristics. Various centrality measures have been used to predict the epidemic outcomes of the nodes, with the basic assumption that more centrally the nodes are located in the network, the greater spreading power they will have[6,13–15]. The well-known centrality measures, and the most frequently used in this field, include: Degree[16], Strength[17], Betweenness[16,18], Closeness[16], Eigenvector[19], PageRank[20] and K-shell[6,7,21–23]. A novel and more effective measure, known as Expected Force[9], has been recently proposed with a particular aim for classifying nodes based on their influence in the network. This measure calculates the possible clusters generated from a fixed number of transmission events, and using entropy as aggregation of the clusters, it generates a node ranking. However, most of the previously proposed centrality measures take into account only the topological features of the network, without explicitly considering the varying strengths (or weights) of the connection strength, which are essential in modelling network diffusion processes. It has been shown that a natural extension to the weighted case is possible[24] for all these measures (as detailed in Supplementary Information) except for K-shell where the

Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong. Correspondence and requests for materials should be addressed to J.L. (email: jiming@comp.hkbu.edu.hk)

ranking is computed from recursive pruning based on the node degrees. To address this issue, a weighted k-shell decomposition[25,26] has been proposed, where the pruning is based on the neighbours' connection strengths.

Another important aspect is concerned with determining the accuracy of the measures, that is to evaluate which measure can be considered as the best super-spreader predictor. In doing so, a susceptible-infected (SI) spreading simulation[10] is typically run starting from each node in the network, so as to obtain a ground truth of the node spreading power. Thus, the problem is translated into that of evaluating which ranking, as calculated based on respective centrality measures, is better correlated with the one generated from simulations. A commonly used method is based on computing correlations (Pearson, Spearman and Kendall tau[27]). However, when the aim becomes the identification of the most influential node, just the top section of the ranking would be important. Besides, different methods, like imprecision function[6] and recognition rate[11], have also been proposed to evaluate a prefixed percentage of the ranking. This allows to have a better comparison because it increases the resolution on the ranking section where the super-spreaders are located. Generally speaking, for many networks, there is no single measure that over performs all the others among different percentages of ranking. For instance, different measures may perform better in different sections of the ranking[15].

Due to the aforementioned observations, it remains a challenge to find which measure, if any, can better identify the super-spreader nodes. This is especially true if considering more realistic constrains, such as a spreading model based on the interaction strengths and an evaluation targeted to the top section of the ranking. To this extend, one can raise the following two questions: (1) Is it possible to find a measure that gives a *consistent* performance among different parts of the ranking for different kinds of networks? (2) Is it possible to find a *corresponding patten* between the best predictors and the characteristics of networks? To the best of our knowledge, there have been no studies in the literature that explicitly and adequately address these questions. In order to answer the aforementioned two questions, in this paper, we investigate a meta-centrality approach in which an aggregation method is used to combine different centrality measures and hence obtain a more robust performance in the super-spreader identification. Furthermore, we conduct a Laplacian spectrum analysis to characterize the networks and to evaluate their correlations with the results obtained from the aggregated measures. Throughout our evaluations, a set of real-word networks covering a broad range of domains has been used.

Results

As mentioned above, there is no single centrality measure that consistently performs as the best predictor. This is because different measures have different objectives, and then based on them, they rank nodes in different ways. In an abstract sense, we may view the ranking computed from a certain centrality measure as that produced by an *expert* that evaluates certain features of a network. Thus, if we could aggregate the rankings (opinions) of different experts, we would be able to improve the final ranking results since each of the aggregated rankings brings its own contribution in the identification. A similar approach has also been adopted in other domains such as meta-search in Web engine[28], biological databases[29], and recommendation systems[30]. In doing so, it is essential to find the best aggregation of different rankings so as to obtain an improvement in the results. There have been some classical aggregation methods, such as Borda count method[31,32], Kemeny-Young method[33,34], and Median rank method. Some of the methods, for instance Kemeny-Young method, demand heavy computation even with a few (e.g., four) different rankings. In this regard, Borda count method requires relatively lighter computation, as it performs a scanning of all the rankings plus a sort. In addition, it also handles rankings with ties well[35], which has been a major problem associated with some of the centrality measure rankings as they assign the same values to different nodes. The original Borda count method considers that every ranking has the same importance. A better performance can be achieved by using just a subset of the rankings or by using a weighted version of the Borda count[32]. Normally, this task involves training data to learn the optimal weighting schema. However, since it is not always possible to have training data, as in our case, other unsupervised models[36] should be considered. In what follows, we first present a novel heuristic method, based on pairwise correlations, for selecting a subset of rankings so as to obtain an improved Borda count aggregation method. Next, we perform a series of experiments on real-network datasets to show the results of our proposed method. Finally, we correlate the topological features of the considered networks with the results as obtained from the improved Borda count aggregation by means of Laplacian spectrum analysis.

The improved Borda count aggregation method. Generally speaking, Borda count method takes as input a set of ordered lists $T=\{\tau_1, \ldots, \tau_n\}$, where each list has the same C items in different orders. Let us denote $\tau_t^{(z)}$ the position of the item z in the list τ_t. Then, the re-ranking value of an item z can be given as follows:

$$B(z) = \sum_{\tau_t \in T} |C| - \tau_t^{(z)} \tag{1}$$

where B is a vector with all the ranking items and $|C|$ is the cardinality of the item set. Thus, the descending reorder of B represents the new aggregated ranking. In this basic form, each rank $\tau_t \in T$ is considered as being equally important, but there could be rankings that have a better accuracy than others. In our present case, the ordered lists correspond to the centrality measure based rankings, whereas the items C are the networks' nodes. As mentioned above, the rankings generated by the centrality measures could be considered as opinions from different experts. Therefore, if a group of experts agree on the same subjects, their opinions could be considered more reliable. In reality, the opinions from experts who do not agree with the trend are still considered reliable, this is because they may still bring in useful alternatives. On the other hand, the situation in which experts are half agreed with each other will be considered unreliable, since their opinions are inconsistent with the mass. Thus, based on this idea of "social" aggregation, we formalize our heuristic method for selecting rankings.The method

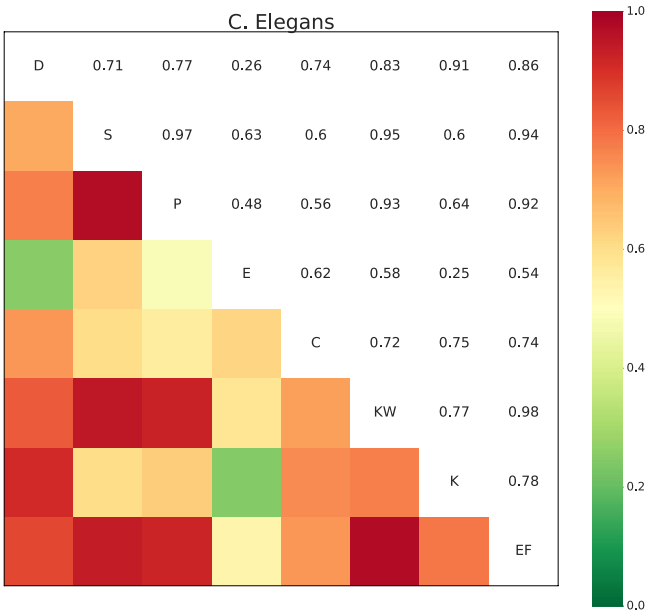

Figure 1. An example of Spearman correlation matrix of C. Elegans network. As a numerical example of the proposed method, we consider just the first row. Thus, we have the sets of $h_1 = \{1, 6, 7, 8\}$, $l_1 = \{1, 4\}$, and $h_1 \cup li = \{1, 4, 6, 7, 8\}$, with $t_b = 0.8$ and $t_s = 0.3$. Then, $E(h_1) = 0.460$, $E(l_1) = 0.505$, and $E(h_1 \cup li) = 0.383$.

proceeds in the following steps: *slicing*, *selection*, and *aggregation*. Let $M = [m_{ij}] \in [0, 1]^{n \times m}$ be the correlation matrix, where $m_{ij} = \rho_{\tau_i \tau_j}$ and ρ represents the Spearman correlation between the rankings τ_i and τ_j. It should be noted that this correlation also enables us to handle any possible disagreement among rankings by assigning it to a negative value. This may happen, for example, when we consider centrality measures such as clustering coefficient[37,38]. Indeed, this centrality assigns lower values to more central nodes, and hence it can lead to negative correlation values[39]. However, in the current work, we will not expect to encounter any large disagreement, since our rankings are based on the selection of those centrality measures that are known to be potential super-spreader identifiers, that is, they exhibit similar monotonic trends by their nature. This is also confirmed in all our experiments in that we have obtained only positive values of correlation except one case with a negative value very close to zero. Generally speaking, in handling such cases, we treat negative correlations similar to the uncorrelated ones (i.e., setting their values to zero).

The details of the three steps are as follows:

Slicing: This step attempts to select the subsets of ranking that are going to be used later in the aggregation step. To do so, we select a series of subsets for each row i in the matrix M. Let be $h_i = \{j | m_{ij} \geq t_b, i \leq j\} \cup \{i\}$ and $l_i = \{j | m_{ij} \leq t_s, i \leq j\} \cup \{i\}$, where t_b and t_s are two positive constant thresholds. Thus, we define the following sets:

$$\begin{aligned} H &= \{h_i\}_{i=1,\ldots,n} \\ L &= \{l_i\}_{i=1,\ldots,n} \\ HL &= \{h_i \cup l_i\}_{i=1,\ldots,n} \end{aligned} \tag{2}$$

where H represents the set containing high correlation subsets, L is the one containing low correlation subsets, and HL is the one having both. Note that different thresholds can create subsets with different cardinalities; we will further discuss about the threshold selection later. Naturally, one can expect that in this way, there could be many subsets, since each row could have different combinations of correlated rankings.

Selection: This step is designed to select the subsets of rankings that contain the most informative attributes. We discard some subsets of the ranking based on their entropy values as follows. Specifically, for each subset of the rankings, $X = H \cup L \cup HL$, we calculate its entropy based on the normalized correlation value. More formally, for each $x_i \in X$, where i is the index of the selected subset row, we compute the following:

$$E(x_i) = -\frac{1}{N} \sum_{j \in x_i} \overline{m_{ij}} \log\left(\overline{m_{ij}}\right) \tag{3}$$

where $\overline{m_{ij}} = \frac{m_{ij}}{\sum_j m_{ij}}$, and $N = |x_i|$ is a scaling factor. This division is used to tune the cardinality of the subsets, in order to pick solutions with a smaller number of rankings. Finally, after the calculation of the entropy, we select

Network	V	E	$<k>$	D	λ_n
Names	1707	9059	10.6	8	1.5000
C. Elegans	297	2148	14.5	5	1.6458
Netsience	379	914	4.8	17	1.7973
FB	1893	13835	14.6	8	1.8667
Advogato	5042	41791	16.6	9	1.8688
Adolescent	2539	10455	8.2	10	1.8696
Geom	3621	9461	5.2	14	1.9513
Astro-ph	14845	119652	16.1	14	1.9550
Hep-th	5818	13644	4.7	19	1.9719
Cond-mat	13861	44619	6.4	18	1.9756
US2013	840	8994	21.4	10	1.9790
US2015	894	8466	18.9	9	1.9899
AS	25241	70669	5.6	10	1.9913
Metro	307	373	2.4	55	1.9941
Rail	2490	4387	3.5	47	1.9971
Coach	1603	2474	3.1	116	1.9987

Table 1. Characteristics of the networks used in the experiments. V is the number of nodes, E is the number of edges, $<k>$ is the average degree of the nodes, D is the network diameter, and λ_n is the largest eigenvalue in the Laplacian matrix.

the subset $x_i = \text{argmax}_{x_i \in X}\{E(x_i)\}$ and, to be more reliable, we also keep the subset with the second highest entropy value. Figure 1 presents a numerical example of this step.

Aggregation: With the above-mentioned *two* selected subsets, we calculate a new score for each node by performing the Borda count aggregation based on Equation (1). Thereafter, with the descending order of the list, we obtain the new ranking.

Experiments on real-world networks. In order to evaluate the effectiveness of our proposed Borda count aggregation-based meta-centrality method, we conduct the super-spreader identification on 16 real-world networks from different domains. Table 1 shows the main features of the 16 networks. For each network, we first perform susceptible-infected simulations so as to establish a ground truth ranking for all the nodes in the network for the purpose of evaluation (See Methods). In doing so, we create a descending ranking I of nodes in which the first node has the largest infection influence. Next, we compute rankings based on centrality measures and use the results as the input of our aggregation method. The centrality measures used in the experiments consist of: Degree (C_D), Strength(C_s), Closeness(C_C), Eigenvector (C_E), PageRank(C_P), K-shell (C_K), weighted K-shell (C_{KW}), and Expected Force (C_{EX}). It should be noted that here apart from K-shell and Degree centrality measures which are unweighted, all the others are calculated based on the edges' weights (See Supplementary Information). We keep the two unweighted measures so as to bring in some contributions in the super-spreader identification from the topological perspectives. For consistency of notation, we let C_* denote a centrality measure, S the node list sorted in a descending according to their ranking scores of C_*, and A the rank generated from the aggregation.

To evaluate the accuracy performances of prediction by the considered centrality measures and by the aggregation, we compare only the top f nodes in the ranking generated from the ground-truth simulations and the one generated by a predictor. This evaluation method, also known as recognition rate[11], can be expressed as follows:

$$R(f) = \frac{\left|I_f \cap S_f\right|}{\left|I_f\right|} \tag{4}$$

where $|*|$ represents the cardinality of a set, I_f the top section of the ranking generated from simulations, and S_f the top section from a centrality measure. In our experiments, we test six value of f, i.e., 0.05, 0.10, 0.15, 0.20, 0.25, and 0.50, and examine the mean and percentile of the recognition rates among them. This allows us to gain a better understanding of the accuracy performance, in particular, concerning the top section of the ranking. For each of the 16 networks, we evaluate all the considered centrality measures along with the two aggregated meta-centrality measures based on the aggregated rankings. Without loss of generality, the parameters of Equation (2) are set as: $t_b = 0.8$ and $t_s = 0.3$.

The above-mentioned two parameters allow for fine-tuning of our method. Here, we made heuristic choices in setting these values in order to select highly correlated as well as uncorrelated rankings. We conducted a series of experiments with values close to the selected ones, and as a result, we found that these values, as a general setting, could identify the correct centrality measures for the aggregation, consistently among all the networks. In other words, we would not make a customized selection of the parameters specifically for each network. Figure 2 shows the mean and percentile of the recognition rates among the considered f range. As shown in Fig. 2, the rankings generated from the aggregation results give the best predictions about super-spreaders for all the networks. In

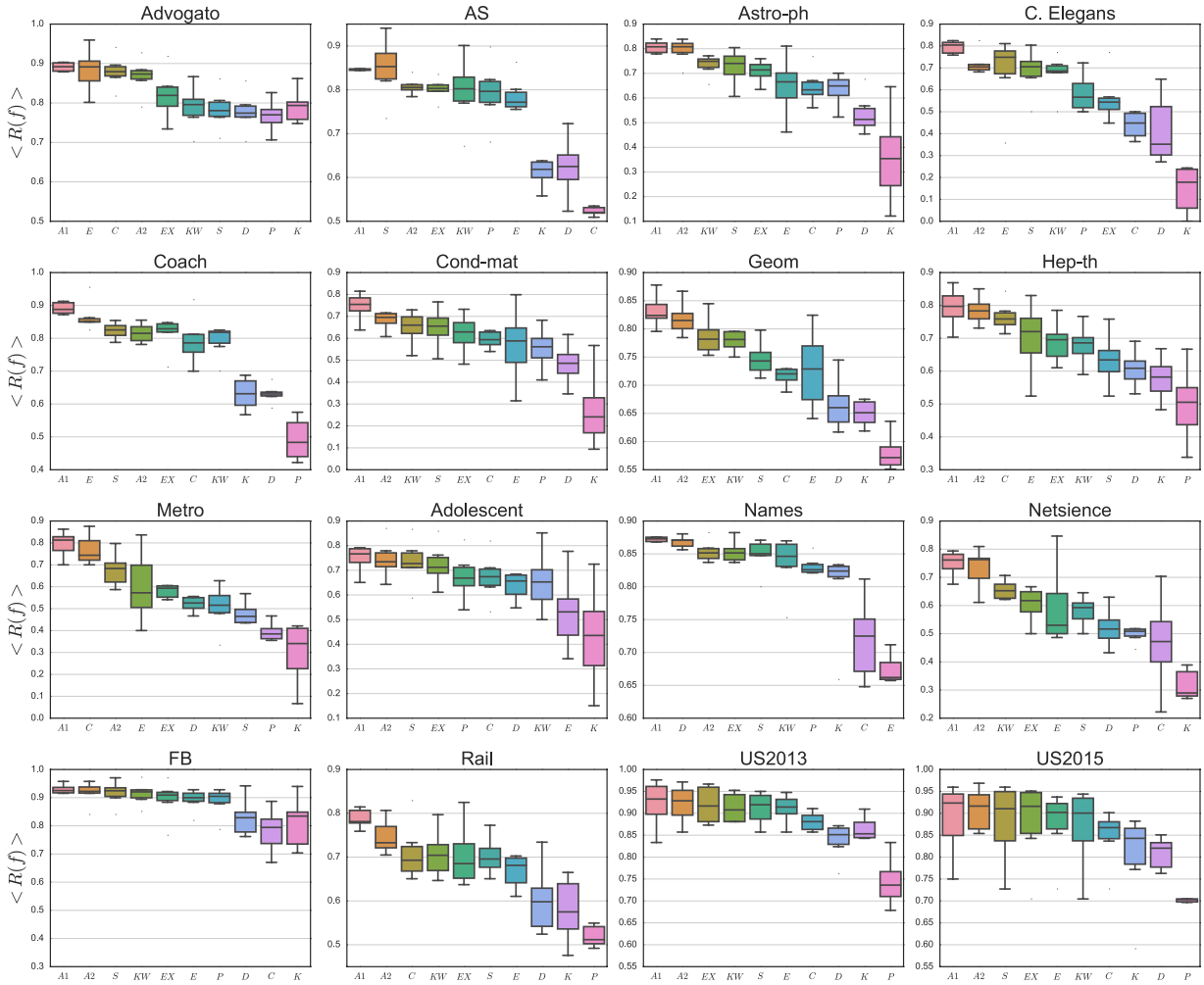

Figure 2. Box-plots showing the recognition rate distribution among different centralities. Each sub-plot shows in a Box-plot the recognition rate variation among *f* values for all the centrality measures and the aggregated ones (as labelled in abbreviation in x-axis). Specifically, the box shows the interquartile range, the segment inside corresponds to the mean value, and two whiskers indicate the maximum and minimum of the range, respectively.

five networks (i.e., C. Elegans, Netsience, Astro-ph, Cond-mat, and Rail), there is a percentage rise of the mean value between 9%–18% as compared to the best single solution. In eight networks (i.e., FB, Geom, Hep-th, Metro, Coach, Advogato, AS, and Adolescent), the percentage rise is between 2–5%. And, in the remaining networks (i.e., US2013, US2015, and Names), the rise is around 1%. Furthermore, in almost all the aggregation results, we observe a decrease in the standard deviation. It should be noted that centrality measures are subject to large variations among different *f*. For instance, Fig. 3 shows the recognition rate values for all the considered centrality measures together with the best aggregation results in four networks. It should also be pointed out that some cases in our experiments, e.g., C. Elegans, Cond-mat, Hep-th, US2013, and US2015, as shown in Fig. 2, where the aggregation selected has the second highest entropy, denoted by A_2 in the subplots, do not lead to the best mean but to a decreased standard deviation value. Therefore, we have used the mean value as the criterion of comparison, and hence the best aggregation from the early-mentioned *selection* step, denoted by A_1 in the subplots, as the proposed solution. For the exact values of the percentages, the best single centrality measure among different *f*, and all the recognition rate plots, check Table S1 and Figures S1-4 in Supplementary Information.

Network characterization based on Laplacian spectrum analysis. Our next task is to examine whether or not there exists a corresponding pattern (or correlation relationship) between centrality measures used for the aggregation and the networks of certain topologies. To achieve this task, we first characterize each of the networks using its Laplacian spectrum. The Laplacian spectrum analysis has been broadly used[40,41] to describe the topology of the network, as it is capable of providing a model of the global topological features of a network, with no explicit reference to its individual nodes. Specifically, in this paper, we consider a Normalized Laplacian Matrix[42], which is defined as $\mathcal{L} = I - D^{1/2}AD^{1/2}$, where A is the adjacency matrix of a network, D is the diagonal matrix of node weighted degrees, and I is the identity matrix. The eigenvalues of $\mathcal{L}$ are in the range of $0 \leq \lambda_1 \leq \ldots \leq \lambda_n \leq 2$,

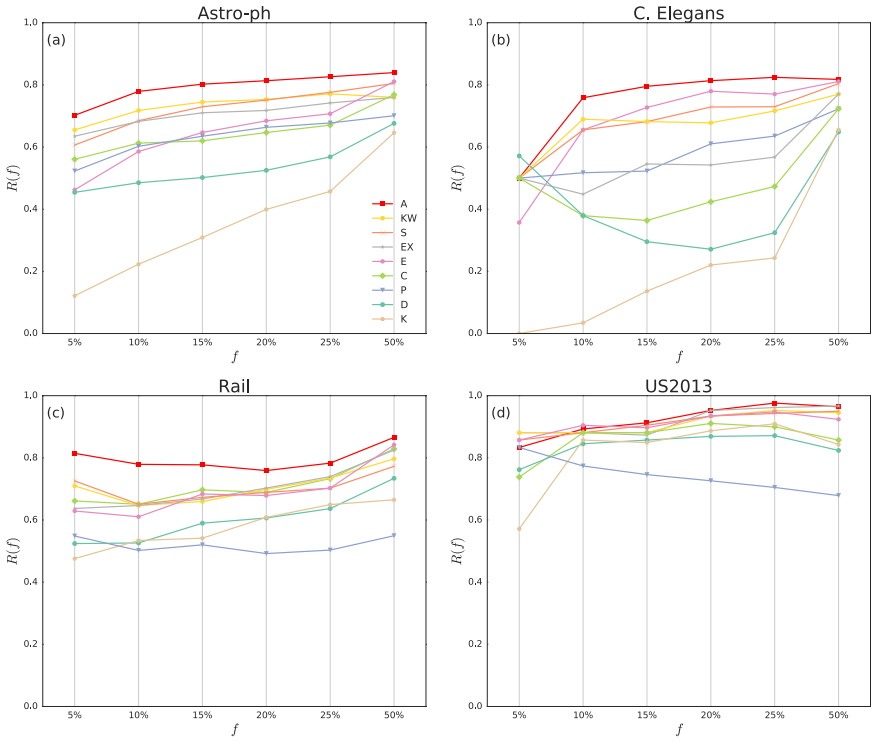

Figure 3. Aggregated solution can better identify super-spreaders as compared to the individual centrality measures. Different values of *f* are given in x-axis, while the corresponding recognition rate changes in y-axis. In this figure, we show the following data-sets: Astro-ph (**a**), C. Elegans (**b**), Rail (**c**), and US2013 (**d**).

where n is the number of nodes in the network[42]. Thus, the Laplacian spectrum of a network refers to the eigenvalue distribution of its Normalized Laplacian Matrix. The Laplacian spectrum characterization provides the information about non-trivial characteristics of a network. There are three well-known characteristics[40]: small eigenvalues imply the presence of community structures[43,44], large eigenvalues reflect the level of bipartiteness[45], and a concentration of eigenvalues near to one indicates the presence of motifs[41]. At the same time, such a characterization allows for comparisons among networks with different sizes and network topologies. For a pair of networks, it is possible to define and compute a pseudo distance between their Laplacian spectra[46] (See Methods) and hence to quantitatively determine their similarity. In our work, having defined a distance between two Laplacian spectra, we use a hierarchy clustering algorithm to group the networks into certain clusters depending on their similarities. This cluster analysis proceeds by considering each observation as a single cluster, and iteratively merging two clusters into a larger one based on a linkage criteria. In our current analysis, we have adopted a complete linkage criteria, which means using the couple with a maximum distance as the distance between two clusters.

Specifically, for each network, we calculate the eigenvalues of its Laplacian matrix. Then, by using the aforementioned methods, we identify four distinct classes of networks. Each identified cluster exhibits certain special network characteristics that influence the identification of super-spreaders. The first cluster that we have identified, i.e., the first sub-plot shown in Fig. 4a, may be referred to as the cluster of transportation networks, consisting of: Coach, Metro, and Rail networks. In this cluster, each of their histograms is almost symmetric, with a concentration of values in the two extremes. This means a high level of bipartiteness[45]. The aggregation in this network cluster uses C_E and C_C as its main components. This centrality couple are also present in other two clusters; indeed all of them have got some large eigenvalues. The second cluster that we have found, i.e., the second sub-plot in Fig. 4b, is composed of networks with the main part of their eigenvalues close to 1. These networks are: Advogato, AS, FB, US2013, and US2015 networks. The Laplacian spectrum of this cluster is typically developed from repeated additions and duplications of nodes and motifs[41]. In this case, two additional components are involved in the aggregation, namely, C_{KW} and C_{EX}. The networks in the third cluster, i.e., the third sub-plot in Fig. 4c, include Astro-ph, Cond-mat, Geom, Hep-th, Adolescent, and Netscience networks. Their eigenvalues tend to concentrate between 1 and 1.5. Similarly to the second cluster, the Laplacian spectra of these networks are also developed from recursive additions of triangle motifs[47]. The aggregation in this network cluster uses a new component C_P. Finally, the forth cluster, shown as the fourth sub-plot in Fig. 4d, is made of two networks, C.Elegans and Names. Although they do not completely share the same spectrum, they are classified together because they do not have significantly large eigenvalues as in the cases of the other network clusters. In this cluster, the component C_D is the common centrality.

In conclusion, we used the Laplacian spectrum to qualitatively characterize the aggregation results. As a result, for each of the four identified network clusters, we revealed a consistent pattern for the clustered network cases. That is, for each network that belongs to the same cluster, the same set of centrality measures in the aggregation

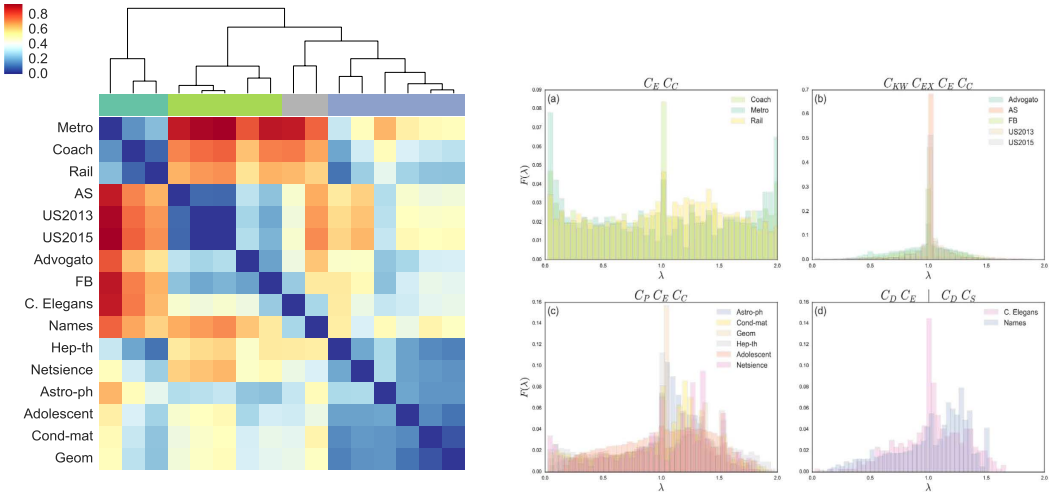

Figure 4. Clusters-map and corresponding spectrum plots. The sub-figure on the left hand side shows the reordered heat map of the spectrum pair distance matrix, with annex dendrogram. The plots on the right hand side present the corresponding histogram plots of the Laplacian spectrum, with eigenvalues in x-axis and their frequencies in y-axis. The four sub-plots correspond to the four identified clusters of the networks, respectively. Note that for the ease of comparison, we have normalized the values of the pair matrix. The title of each sub-plot indicates the set of centrality measures used for the aggregation.

could be found. It would be desirable to make an even stronger conclusion on the connection between network topologies and centralities (or an aggregation of them) that optimally identifies super-spreaders. However, due to the lack of sufficient network samples, it remains to be a future task to quantitatively establish and/or analyze such a connection.

Discussion

Our work has shown that the aggregation of individual centrality measures results in an excellent predictor for super-spreaders, since it brings in contributions from each of individual ones so as to improve the identification accuracy. Most of the considered centrality measures work in a linear time, but not all. In fact, the closeness centrality does not work well for large networks. To overcome this problem, approximated version of this centrality has been proposed and can be used[48].

The next important issue to highlight is about the simulation parameterization. Since our aim is to identify the most influential spreaders, the parameter α has been set equal to one, in order to perform a simulation that reflects the importance of the weights in the infection propagation. Indeed, if α is set to a large value, most of the nodes result in having similar outcomes, and thus we will not have a sufficient resolution to distinguish the importance of the nodes. Therefore, we have decided to fix the parameter α based on the following two considerations: First, we would like to bind the infection propagation only to the edge weights so as to reflect as much as possible the real importance of each node. Second, after a series of experiments, in each network, we have noticed that using different values of α would only increase the spreading power of every single node, while not affecting the final comparison of the nodes. Indeed, the model[49] that we have implemented is made to bind the infection propagation to the edge weights (See Methods). Different methods have been proposed[49–51] to achieve this task, but they were tailor made for particular applications and not suited to the problem of super-spreader identification.

Last but not the least, it should be pointed out that in our current work, we have chosen the SI model as the main spreading model. This is due to the consideration that it fits well with the task of super-spreader identification, in which infected nodes can spread freely until the entire network is fully covered. Nevertheless, it would also be desirable and interesting to examine the results of this work by adopting another commonly used model, i.e., the Susceptible-Infected-Recovered (SIR) model. When using the SIR model, the average numbers of Recovered nodes[6] at the end of the epidemic spreading could be used for the ground truth ranking. In this study, in order to make sure that our method can also be applied to such a model, we have conducted several experiments by using our aggregation model on SIR simulations. The aggregated solution still gives the best results among all the networks. Interested readers are referred to more details in the Supplementary Information.

Methods

In what follow, $G(V, E)$ denotes an undirected, connected weighted network, where V represents the set of nodes, and $E = V \times V$ the set of edges, and w_{ij} represents the weight of the edge $e(v_i, v_j)$. Let us denote with A and W the adjacency matrices of the network G, where $A_{ij} = 1$ represents the edge $e(v_i, v_j)$ and $W_{ij} = w_{ij}$ represents the weight of the connection.

Spreading models and comparisons. As a measure of spreading power of a node, without loss of generality, here we describe a susceptible-infected (SI) simulation model. In the SI model, at the beginning, all the nodes are

in the susceptible state (S) except the one that is in the infected state (I). At each time step of the simulation, the infected nodes will spread the infection to their neighbors, depending on the weights of the connecting edges. The simulation stops when all the network nodes are covered (labeled as I). Different models have been proposed to characterize an infection propagation process based on the connections' strength[49–51]; in this paper, we adopt the most commonly used model[49]. In this model, the probability that a node i becomes infected at the time t is given by $\delta_i(t) = 1 - \prod_{j \in N_i(t)} (1 - \delta_{ij})$, where $\delta_{ij} = \left(\frac{w_{ij}}{w_M}\right)^{\alpha}$, α is a positive constant that describes the infection power and allows to tune the power and the speed of the infection, $N_i(t)$ are the infected neighbors at the time t, and w_M is the largest value of w_{ij} in the network.

To characterize the spreading power of each node, we run a batch of 100 simulations for each seed. In doing so, we record the infection ratio at each time step of a single simulation run in a set, until it has covered all the network nodes (note that the size of the set may vary among different simulation runs). Then, for each generated set, we take the average of its recorded values, and place it in a list L_{sim} (each value corresponds to a single simulation run). Finally, we calculate the average of the list L_{sim} previously generated, so as to obtain a single index that represents the spreading power of a node (more details can be found in Supplementary Information). Another way to characterize the spreading power of a node is to consider the average time of full coverage among different simulations. This method has been frequently used[9], however with the above discussed method we can achieve a better characterization, without losing track of the average time of full coverage. A further discussion can be found in Supplementary Information under the SI evaluation section.

Laplacian spectrum. The normalized Laplacian matrix of a weighted network G is defined as $\mathcal{L} = I - D^{1/2}AD^{1/2}$, where A is the adjacency matrix, D is the diagonal matrix of node weighted degrees, and I the is the identity matrix. Therefore, for each node $u, v \in V$, we have:

$$\mathcal{L}(u, v) = \begin{cases} 1 - \frac{w_{uv}}{d_v} & \text{if u = v, and } d_j \neq 0, \\ -\frac{w_{uv}}{\sqrt{d_u d_v}} & \text{if u and v are adjacent,} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where $d_u = \sum_v w_{uv}$. The eigenvalues of $\mathcal{L}$ define the spectrum of G, and they are in the range of $0 \leq \lambda_1 \leq \ldots \leq \lambda_n \leq 2$, where $n = |V|$[42]. For each network N_i, we compute the corresponding spectrum of the Laplacian matrix $\mathcal{L}_i$ and we define the corresponding eigenvalue histogram as $Hist(\mathcal{L}_i) = \{(b_{i1}, v_{i1}), \ldots, (b_{ij}, v_{ij})\}$, where b_{ij} represents the $j-th$ bin and v_{ij} its frequency. In our analysis, we fix the bin number to 200, such that every bin has a length of 0.01. Since every histogram has got the same number of bins, we can introduce a metric to formally quantify the distance between two spectra. Different metrics have been proposed to achieve this task, either by using a Gaussian kernel estimation[52] or based on an Euclidean distance between the entire spectrum[44]. Instead, in this paper, we use the Earth mover's distance (EMD), that is equal to Wasserstein distance, as the metric. This distance falls into the category of cross-bin measures as used to calculate distances between histograms. Computing EMD for multidimensional histograms generally requires the introduction of an optimization framework[53]. Since here we deal with a mono-dimensional histogram, the EMD becomes the distance between the cumulative histograms. More formally, the distance between two spectra $\mathcal{L}_z$ and $\mathcal{L}_t$ is defined as: $d(\mathcal{L}_z, \mathcal{L}_t) = \sum_{i=1}^{m} |CDF_i(Hist(\mathcal{L}_z)) - CDF_i(Hist(\mathcal{L}_t)|$, where CDF_i is the cumulative histogram defined as $CDF_i(Hist(\mathcal{L}_z)) = \sum_{k=1}^{i} v_{zk}$. We choose this approach because it is simple and straightforward, and has been proven to be a pseudo metric as well as a graph spectrum distance[46].

Data Sets. In this work, all the networks are considered undirected, weighted, and connected. A brief description of the networks used is given as follows: (1) Names (nouns of the King James bible and their occurrences)[54], (2) C. Elegans (weighted network representing a neural network)[55], (3) Netscience (co-authorship network of scientists working on network theory and experiments)[56] (4) FB (online community of students at University of California)[57], (5) Advogato (online community for developers)[54], (6) Adolescent (created from a survey that took place in 1994/1995, friendship choices by students)[54], (7) Geom (authors collaboration network in computational geometry)[58], (8) Astro-ph (co-authorships between scientists posting preprints on the Astrophysics E-Print Archive)[59], (9) Hep-th (co-authorships between scientists posting preprints on the High-Energy Theory E-Print Archive)[59], (10) Cond-mat (co-authorships between scientists posting preprints on the Condensed Matter E-Print Archive)[59], (11) and (12) US air passengers 2013 and 2015[60], (13) AS (snapshot of Autonomous System network, where the weights representing the visit-counts in the period of time)[61], and (14), (15) and (16) Metro, Coach and Rail respectively (network map of UK transportation system, where the edge weights are the average minutes of travel)[62,63].

References

1. Valente, T. W. *Network models of the diffusion of innovations*, vol. 2 (Hampton Press Cresskill, NJ, 1995).
2. Watts, D. J., Peretti, J. & Frumin, M. *Viral marketing for the real world* (Harvard Business School Publishing, 2007).
3. Christakis, N. A. & Fowler, J. H. Social network sensors for early detection of contagious outbreaks. *PloS one* **5,** e12948 (2010).
4. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438,** 355–359 (2005).
5. Gallos, L. K., Liljeros, F., Argyrakis, P., Bunde, A. & Havlin, S. Improving immunization strategies. *Physical Review E* **75,** 045104 (2007).
6. Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nature Physics* **6,** 888–893 (2010).

7. Liu, Y., Tang, M., Zhou, T. & Do, Y. Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition. *Scientific Reports* **5,** 9602 (2015).
8. Liu, Y., Tang, M., Zhou, T. & Do, Y. Improving the accuracy of the k-shell method by removing redundant links: From a perspective of spreading dynamics. *Scientific Reports* **5,** 13172 (2015).
9. Lawyer, G. Understanding the influence of all nodes in a network. *Scientific Reports* **5,** 8665 (2015).
10. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Reviews of Modern Physics* **87,** 925 (2015).
11. Pei, S., Muchnik, L., Andrade Jr, J. S., Zheng, Z. & Makse, H. A. Searching for superspreaders of information in real-world social media. *Scientific Reports* **4,** 5547 (2014).
12. Mela, S. & Whitworth, D. E. The fist bump: A more hygienic alternative to the handshake. *American Journal of Infection Control* **42,** 916–917 (2014).
13. Canright, G. S. & Engø-Monsen, K. Spreading on networks: a topographic view. *Complexus* **3,** 131–146 (2006).
14. Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C. & Zhou, T. Identifying influential nodes in complex networks. *Physica a: Statistical mechanics and its applications* **391,** 1777–1787 (2012).
15. Macdonald, B., Shakarian, P., Howard, N. & Moores, G. Spreaders in the network sir model: An empirical study. *arXiv preprint arXiv:1208.4269* (2012).
16. Freeman, L. C. Centrality in social networks conceptual clarification. *Social networks* **1,** 215–239 (1979).
17. Barthélemy, M., Barrat, A., Pastor-Satorras, R. & Vespignani, A. Characterization and modeling of weighted networks. *Physica a: Statistical mechanics and its applications* **346,** 34–43 (2005).
18. Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* 35–41 (1977).
19. Bonacich, P. & Lloyd, P. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks* **23,** 191–201 (2001).
20. Brin, S. & Page, L. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks* **56,** 3825–3833 (2012).
21. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. K-core organization of complex networks. *Physical Review Letters* **96,** 040601 (2006).
22. Alvarez-Hamelin, J. I., Dall'Asta, L., Barrat, A. & Vespignani, A. How the k-core decomposition helps in understanding the internet topology. *Proceedings of ISMA Workshop on the Internet Topology*, San Diego, California, May 10–12, 2006, vol. 1 (CAIDA, 2006).
23. Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y. & Shir, E. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences* **104,** 11150–11154 (2007).
24. Opsahl, T., Agneessens, F. & Skvoretz, J. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks* **32,** 245–251 (2010).
25. Garas, A., Schweitzer, F. & Havlin, S. S-core network decomposition: A generalization of k-core analysis to weighted networks. *New Journal of Physics* **14,** 083030 (2012).
26. Eidsaa, M. & Almaas, E. S-core network decomposition: A generalization of k-core analysis to weighted networks. *Physical Review E* **88,** 062819 (2013).
27. Kendall, M. G. A new measure of rank correlation. *Biometrika* 81–93 (1938).
28. Dwork, C., Kumar, R., Naor, M. & Sivakumar, D. Rank aggregation methods for the web. *Proceedings of the 10th International Conference on World Wide Web,* Hong Kong, Hong Kong, 613–622 (ACM, 2001).
29. DeConde, R. P. *et al.* Combining results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology*, **5,** 1544–6115 (2006).
30. Baskin, J. P. & Krishnamurthi, S. Preference aggregation in group recommender systems for committee decision-making. *Proceedings of the 3th ACM Conference on Recommender Systems*. New York, USA, 337–340 (ACM, 2009).
31. De Borda, J. C. *Mémoire sur les élections au scrutin*. Histoire de l'Academie Royale des Sciences (1781).
32. Aslam, J. A. & Montague, M. Models for metasearch. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New Orleans, USA, 276–284 (ACM, 2001).
33. Kemeny, J. G. Mathematics without numbers. *Daedalus* **88,** 577–591 (1959).
34. Young, H. P. & Levenglick, A. A consistent extension of condorcet's election principle. *SIAM Journal on Applied Mathematics* **35,** 285–300 (1978).
35. Brancotte, B. *et al.* Rank aggregation with ties: Experiments and analysis. *Proceedings of the VLDB Endowment,* July. **8,** 1202–1213, doi: 10.14778/2809974.2809982 (2015).
36. Klementiev, A., Roth, D. & Small, K. An unsupervised learning algorithm for rank aggregation. *Proceedings of the 18th European Conference on Machine Learning,* Warsaw, Poland, Sept 17–21, 2007, 616–623 (Springer, 2007).
37. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. Complex networks: Structure and dynamics. *Physics Reports* **424,** 175–308 (2006).
38. Newman, M. *Networks: An introduction* (Oxford university press, 2010).
39. de Arruda, G. F. *et al.* Role of centrality for the identification of influential spreaders in complex networks. *Physical Review E* **90,** 032812 (2014).
40. de Lange, S. C., de Reus, M. A. & van den Heuvel, M. P. The laplacian spectrum of neural networks. *Frontiers in Computational Neuroscience*. **7,** 189 (2013).
41. Banerjee, A. & Jost, J. Spectral characterization of network structures and dynamics. *Dynamics on and of Complex Networks*, 117–132 (Springer, 2009).
42. Chung, F. R. *Spectral Graph Theory*. vol. 92 (American Mathematical Society, 1997).
43. Cetinkaya, E. K., Alenazi, M. J., Rohrer, J. P. & Sterbenz, J. P. Topology connectivity analysis of internet infrastructure using graph spectra. *Proceedings of 4th International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT)*. St. Petersburg, Russia, Oct 3–5, 2012, 752–758 (IEEE, 2012).
44. Wilson, R. C. & Zhu, P. A study of graph spectra for comparing graphs and trees. *Pattern Recognition* **41,** 2833–2841 (2008).
45. Bauer, F. & Jost, J. Bipartite and neighborhood graphs and the spectrum of the normalized graph laplacian. arXiv preprint arXiv:*0910.3118* (2009).
46. Gu, J., Hua, B. & Liu, S. Spectral distances on graphs. *Discrete Applied Mathematics* **190,** 56–74 (2015).
47. Banerjee, A. & Jost, J. On the spectrum of the normalized graph laplacian. *Linear Algebra and its Applications* **428,** 3015–3022 (2008).
48. Okamoto, K., Chen, W. & Li, X.-Y. Ranking of closeness centrality for large-scale social networks. *Frontiers in Algorithmics*. 186–195 (Springer, 2008).
49. Gang, Y., Tao, Z., Jie, W., Zhong-Qian, F. & Bing-Hong, W. Epidemic spread in weighted scale-free networks. *Chinese Physics Letters* **22,** 510 (2005).
50. Kamp, C., Moslonka-Lefebvre, M. & Alizon, S. Epidemic spread on weighted networks. *PLoS Computational Biology* **9,** e1003352 (2013).
51. Sun, Y., Liu, C., Zhang, C.-X. & Zhang, Z.-K. Epidemic spreading on weighted complex networks. *Physics Letters A* **378,** 635–640 (2014).
52. Gu, J., Jost, J., Liu, S. & Stadler, P. F. Spectral classes of regular, random, and empirical graphs. *Linear Algebra and its Applications* **489,** 30–49 (2016).

53. Rubner, Y., Tomasi, C. & Guibas, L. J. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision* **40,** 99–121 (2000).
54. Kunegis, J. Konect: the koblenz network collection. *Proceedings of the 22nd International Conference on World Wide Web,* Rio de Janeiro, Brazil, 2013, 1343–1350 (ACM, 2013).
55. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world'networks. *Nature* **393,** 440–442 (1998).
56. Newman, M. E. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* **74,** 036104 (2006).
57. Panzarasa, P., Opsahl, T. & Carley, K. M. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology* **60,** 911–932 (2009).
58. Batagelj, V. & Mrvar, A. *Pajek datasets. collaboration network in computational geometry* Available at: http://vlado.fmf.uni-lj.si/pub/networks/data/collab/geom.htm. (Accessed: 20/02/2016) (2006).
59. Newman, M. E. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* **98,** 404–409 (2001).
60. Office of the Assistant Secretary for Research and Technology (OST-R) U.S. *Department of Transportation* (*US DOT*) *Airline passengers, flights, freight and other air traffic data* Available at: http://www.transtats.bts.gov/. (Accessed: 04/01/2016) (2014).
61. Shavitt, Y. & Shir, E. Dimes: Let the internet measure itself. *ACM SIGCOMM Computer Communication Review* **35,** 71–74 (2005).
62. Gallotti, R. & Barthelemy, M. The multilayer temporal network of public transport in Great Britain. *Scientific Data* **2,** 140056; doi: 10.1038/sdata.2014.56 (2015).
63. Gallotti, R. & Barthelemy, M. *The multilayer temporal network of public transport in Great Britain*. Available at: http://dx.doi.org/10.5061/dryad.pc8m3. (Accessed: 04/12/2015) (2015).

Acknowledgements

The authors would like to acknowledge support from Hong Kong Research Grants Council (HKBU211212 & HKBU12202415).

Author Contributions

Conceived and designed the experiments: J.L. A.M. Performed the experiments: A.M. Analysed the data and experimental results: J.L. A.M. Wrote and revised the paper: J.L. A.M. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Madotto, A and Liu, J. Super-Spreader Identification Using Meta-Centrality. *Sci. Rep.* **6**, 38994; doi: 10.1038/srep38994 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.