



LETTER

## Influence maximization in noisy networks

To cite this article: irag Erkol *et al* 2018 *EPL* **123** 58007

View the [article online](#) for updates and enhancements.

# Influence maximization in noisy networks

ŞIRAG ERKOL<sup>1</sup>, ALI FAQEEH<sup>1,2</sup> and FILIPPO RADICCHI<sup>1</sup>

<sup>1</sup> *Center for Complex Networks and Systems Research, School of Informatics, Computing, and Engineering, Indiana University - Bloomington, IN 47408, USA*

<sup>2</sup> *MACSI, Department of Mathematics & Statistics, University of Limerick - Limerick, Ireland*

received 6 March 2018; accepted in final form 6 September 2018  
published online 5 October 2018

PACS 87.23.Ge – Dynamics of social systems  
PACS 89.75.Hc – Networks and genealogical trees

**Abstract** – We consider the problem of identifying the most influential nodes for a spreading process on a network when prior knowledge about structure and dynamics of the system is incomplete or erroneous. Specifically, we perform a numerical analysis where the set of top spreaders is determined on the basis of prior information that is artificially altered by a certain level of noise. We then measure the optimality of the chosen set by measuring its spreading impact in the true system. Whereas we find that the identification of top spreaders is optimal when prior knowledge is complete and free of mistakes, we also find that the quality of the top spreaders identified using noisy information does not necessarily decrease as the noise level increases. For instance, we show that it is generally possible to compensate for erroneous information about dynamical parameters by adding synthetic errors in the structure of the network. Further, we show that, in some dynamical regimes, even completely losing prior knowledge on network structure may be better than relying on certain but incomplete information.

Copyright © EPLA, 2018

**Introduction.** – In a social network where an opinion diffuses according to an irreversible spreading process, a fundamentally important role for the ultimate success of the opinion is played by the nodes that act as initiators or seeds of the spreading process. For example, the popularity of memes in social media is often determined by just a few early adopters [1]. Given the high sensitivity of the outcome of spreading processes to initial configurations, a very interesting problem regards the identification of the initial configuration, among the many possible, that maximizes the extent of diffusion. The problem is traditionally named as influence maximization. It has been first considered by Domingos and Richardson [2], and slightly later generalized by Kempe *et al.* [3]. Roughly speaking, influence maximization consists in an optimization problem based on a few assumptions and subjected to one constraint. The function that one wants to maximize is the size of the outbreak, *i.e.*, the number of nodes that will end up acquiring the opinion that is diffusing in the system. The assumptions in the formulation of the problem regard the structure of the network and the type of spreading that is taking place on the network. This information is generally assumed as a prior knowledge, and it is actively used for finding solutions to the optimization problem. The

only constraint used in the optimization problem is the number of seeds. Only initial configurations consisting of a given number of active nodes are considered as potential solutions to the optimization problem.

As most optimization problems, influence maximization is NP-complete [3]. Exact solutions can be found only in very small networks. Suboptimal solutions can be achieved with approximated or greedy optimization techniques [3,4]. These approaches are generally effective, but they are designed for the analysis of small to medium networks. The identification of influential spreaders in large networks is allowed only through the use of heuristic techniques where dynamics is *de facto* neglected, and the solution to the optimization problem is approximated relying on network centrality metrics [5–11]. This approach finds its rationale in interpreting the high sensitivity of the outcome of a spreading process to the initial conditions as a consequence of the heterogeneity of the underlying network. However, geometry alone is not enough to provide a sufficiently accurate description of the state of a dynamical system running on a network [12]. The identity of the most influential nodes in a network generally changes from type to type of spreading process, and, even for the same type of process, it may depend on its dynamical regime [13].

Most of the studies we mentioned above rely on one strong assumption: prior knowledge of system structure and dynamics is complete and free of errors. When dealing with a real application of influence maximization, we should however recognize that this assumption is at least optimistic. The presence/absence of a connection in a social network is generally established from the result of some experimental observation, and it is therefore potentially affected by experimental errors [14]. Similarly, we may be aware of the type of process that drives spreading, but we may be unsure about the exact value of the rates at which spreading occurs. There are techniques that allow to accurately estimate spreading rates from empirical observations of spreading events [15]. However, these techniques rely on the assumption that structural information is complete and free of mistakes. Further, in influence maximization, one aims at controlling the fate of a future or ongoing spreading process, so posterior estimates of the rates are not very helpful.

Several previous studies have considered the reliability of network centrality metrics when computed from noisy or incomplete structural information [16,17]. In the context of influence maximization, we are aware of previous tests of robustness of some centrality metrics in noisy structural data [5]. However, to the best of our knowledge, there are no previous studies that attempted to understand how incomplete or erroneous information about both structure and dynamics affects our ability to solve the problem of influence maximization. Please note that we may naively expect that noise does not dramatically modify the overall trend of a geometric centrality metric, as was shown in ref. [5]. However, the distortions that noise can create in the solutions of an optimization problem such as influence maximization are far less predictable. The current paper aims at filling this gap of knowledge.

## Methods. –

*Network structure.* We assume that the network where the spreading process takes place and where we aim at solving the problem of influence maximization is given by  $N$  total nodes and  $M$  edges. The network is unweighted and undirected. Structural information about the network is fully specified by the adjacency matrix  $A$ , whose generic element  $A_{ij} = A_{ji} = 1$  if nodes  $i$  and  $j$  are connected, and  $A_{ij} = A_{ji} = 0$ , otherwise. Note that, in the spreading process, only the state associated to the nodes of the network can change. Edges do not have states that evolve in time, but serve only as static media for spreading.

*Spreading dynamics.* In this paper, we focus our attention on a spreading model that is very popular in studies about the identification of influential spreaders in networks: the Independent Cascade Model (ICM) [3]. The ICM is very similar to the traditional Susceptible-Infected-Recovered (SIR) model [18]. During ICM dynamics, nodes in the network can be found in three different states: S, I, or R. Generally, in an initial configuration of the dynamics, all nodes are set in state S, except for a subset of seeds

$Q$  that are set in state I. At each discrete stage of the dynamics, two rules are applied in sequence: i) every node in state I infects, with probability  $p$ , each of its neighbors in state S; ii) all nodes in state I, that attempted to infect their neighbors at step i), recover and change their state from I to R. Rules i) and ii) are iterated until infected nodes are no longer present in the network. The size of the outbreak is given by the total number of nodes that are in state R at the end of the dynamics. This number is a stochastic variable that may change its value from instance to instance of the model. Observed values depend on the network structure encoded by the adjacency matrix  $A$ , the value of the spreading probability  $p$ , and the set of seeds  $Q$  that initiated the spreading process. In our numerical study, we measure the spreading intensity for given  $A$ ,  $p$ , and  $Q$  in terms of the average value of outbreak size, namely  $O$ , over 20 independent simulations of the spreading process.

*Influence maximization.* Solving the problem of influence maximization for a spreading process on a network means finding the set of seeds that maximizes the average value of the outbreak size  $O$ . The maximization is performed for a fixed size  $|Q|$  of the set of seeds. The optimization relies on prior knowledge about the structure of the network and the dynamical rules of the spreading process. Information on the structure is provided by the adjacency matrix  $A$ . Information about the dynamics consists in knowing that spreading is regulated by the ICM and the value of the probability of spreading is  $p$ . Even with full knowledge of system structure and dynamics, identifying the set of optimal seeds is a NP-complete problem, and thus exact solutions are achievable only in extremely small networks [3]. Greedy optimization, however, allows to provide suboptimal solutions that are granted to be within 63% of the optimum [3]. This is the consequence of the fact that the size of the outbreak  $O$  is a submodular function of the set of seeds [3,19]. According to greedy optimization, the set of influential spreaders  $Q$  is constructed sequentially by adding one node at a time. At every stage, the best node to be added to the set of seeds is the one that leads to the maximal value of  $O$ , in a spreading process initiated by all nodes that are already part of the seed set plus the node under consideration. In the original version of the algorithm by Kempe *et al.* [3], greedy steps rely on direct simulations of the dynamical process. This algorithm is very general, and can be used for other types of dynamical models, not just the ICM. For the specific case of the ICM, the greedy algorithm can be further speeded up by taking advantage of the mapping between static properties of the SIR and bond percolation [20]. The mapping allows to use configurations of the percolation model to infer information about final configurations arising from ICM dynamics. Each node is associated with a score whose value is proportional to the average size of independent clusters (*i.e.*, clusters that do not contain previously identified seed nodes) which the node belongs to. Now the set  $Q$  is constructed by adding the nodes to

it, one by one, starting from the nodes with the highest score. This method was designed, implemented and validated by Chen *et al.* [4]. The results of the current paper are based on our re-implementation of the algorithm by Chen *et al.* We rely on  $R = 1000$  independent simulations of the bond percolation model to compute the scores associated with the nodes. Please note that as long as  $R$  is a finite number, scores associated with nodes are subjected to finite-size fluctuations. As a consequence, the solution to the optimization problem provided by the algorithm by Chen *et al.* may depend on the specific set of bond percolation simulations considered. As the results of the Supplementary material `Supplementarymaterial.pdf` (SM) show, we verified that, while the identity of the nodes in the solutions provided by the algorithm is not always the same, the average sizes  $O$  of the outbreak associated with those solutions are very similar. This finding suggests that the exact detection of the set of top spreaders is statistically irrelevant for the outcome of the spreading process, in the sense that there are many nearly optimal solutions to the problem of influence maximization.

*Modeling errors in system dynamics and structure.*

The process of selection of top spreaders relies on prior knowledge about the structure of the underlying network and the details of the dynamical process. This means that the set  $Q$  depends on the information at our disposal regarding the structure of the network, *i.e.*, the adjacency matrix  $A$ .  $Q$  also depends on our prior knowledge about the dynamical process that is taking place on the network, that is the ICM model with spreading probability equal to  $p$ . It is common practice to assume prior information complete and exact. This is equivalent to assuming that the inputs  $A$  and  $p$  of the algorithm for the identification of top spreaders are equal to their true values, namely  $A_{true}$  and  $p_{true}$ , respectively. However, in practical situations, this may not be the case. Prior knowledge may be affected by errors, so that the actual information used to solve the problem of influence maximization is given by  $A_{err}$  and  $p_{err}$ , respectively. There are potentially many different ways to model errors that deteriorate dynamical and structural information of the system. Here, we opt for simple, yet realistic, models.

Errors that affect our prior knowledge of the spreading dynamics are simply obtained by setting  $p_{err} \neq p_{true}$ . Essentially, we assume to know that spreading occurs following the rules of the ICM, but we pretend that we do not know the exact value of the spreading probability. Instead of using raw values for the spreading probabilities, we rescale them as  $\phi_{true} = p_{true}/p_c$  and  $\phi_{err} = p_{err}/p_c$ , where  $p_c$  is the critical value of the spreading probability for the ICM model on the true network  $A_{true}$ . This transformation is used only to simplify the presentation of the results. It allows us in fact to use the same reference value for all networks to distinguish between different dynamical regimes of spreading. The critical value of the spreading probability is computed directly from numerical simulations where ICM is initiated from a randomly

selected seed node [21]. Depending on whether  $\phi_{true}$  is larger, equal to or smaller than one, we say that the network is respectively in the supercritical, critical or subcritical regime of spreading. Similarly, the value of  $\phi_{err}$  tells us what type of dynamical regime is hypothesized for the selection of influential spreaders. From previous studies, we know that the identity of the nodes that populate the sets of top spreaders is highly dependent on the regime of spreading [13]. We expect therefore that the error in the estimation of the spreading probability may strongly affect our ability to properly predict top spreaders.

Errors in the structure of the network are generated artificially according to a model similar to the one considered in ref. [14]. The total number of nodes is unaffected by noise, so that we can indicate it as  $N$ . Errors happen at the level of pairs of nodes. This means that the total number of edges  $M_{err}$  in the altered structure generally differs from  $M_{true}$ , *i.e.*, the number of edges in the true network. We consider two potential sources of errors. The first source is responsible for making true edges disappear from our prior knowledge of the network. Specifically, given the true adjacency matrix  $A_{true}$ , every pair of connected nodes is disconnected with probability  $0 \leq \epsilon_{del} \leq 1$  in  $A_{err}$ . The number of true edges that are deleted equals zero for  $\epsilon_{del} = 0$ , and equals  $M_{true}$  for  $\epsilon_{del} = 1$ . The second source of error generates false edges. Every pair of nodes that is not connected according to  $A_{true}$  appears as connected in  $A_{err}$  with probability  $\epsilon_{add} M_{true}/[N(N-1)/2 - M_{true}]$ , where  $0 \leq \epsilon_{add} \leq 1$ . For  $\epsilon_{add} = 0$ , no false edges are added to the true network. For  $\epsilon_{add} = 1$  instead, the expected total number of false edges equals  $M_{true}$ . The definition of the noise parameters  $\epsilon_{del}$  and  $\epsilon_{add}$  are such that both parameters are confined in the range  $[0, 1]$ , and their maximum values correspond to an expected alteration (*i.e.*, deletion or addition) of 100% of the true number of edges.

Please note that removing true edges is generally not the inverse operation of adding false edges. For instance, even the addition of a very small number of edges among pairs of non-connected nodes is able to decrease substantially the average path length of graphs that do not originally satisfy the small-world property [22]. For example, in networks with strong spatial embedding, as some of those we analyze in this paper, random edges likely behave as shortcuts between spatially far regions of the system. On the other hand, the removal of a small fraction of true edges does not change dramatically the average path length of the graph. Also, we do not expect the two sources of structural errors to be equally likely in real networks. Their proportion may depend strongly on the type of network considered, and on the way the network is actually constructed from empirical observations [14]. As the two sources of structural errors cannot be treated on the same footing, in our analysis we always consider them separately, *i.e.*, we always work with the condition  $\epsilon_{del}\epsilon_{add} = 0$  satisfied.

We stress that our choice for the noise model affecting prior structural information is heavily inspired by ref. [14].

We find the model simple enough, yet able to naturally describe sources of uncertainty in empirically constructed social networks. Alternative models of structural noise could be considered and studied using the same exact methods as those described here. For example, a model consisting in shuffling true edges with a certain probability would provide a way to introduce structural noise without altering the degree of the nodes in the network. In general, we believe that the choice of the noise model should depend on the specific question that one wants to address, or the specific system that one is considering. The set of questions that we are considering here can be fully addressed by the particular choice we made.

*Measuring performance.* Given the inputs  $A_{err}$  and  $\phi_{err}$ , we make use of the algorithm by Chen *et al.* to identify the set  $Q_{err}$  of top spreaders. As the algorithm for the identification of top spreaders is based on a finite number of numerical simulations and the output of the algorithm is subjected to finite-size fluctuations, we apply the algorithm  $V = 10$  times to find  $V$  potentially different sets  $Q_{err}$ . For each of them, we use numerical simulations of the ICM model relying on  $A_{true}$  and  $p_{true}$  to evaluate their performance. In particular, as the algorithm by Chen *et al.* naturally ranks nodes according to the order in which they are added to the set of top spreaders, we explicitly use this information to quantify the performance of the set  $Q_{err}$  as follows. The size of the set  $Q_{err}$  is indicated with  $|Q_{err}|$ . Nodes have been added to the set according to the sequence  $q_1, q_2, \dots, q_{|Q_{err}|}$ . Define  $Q_{err}^{(r)} = \cup_{v=1}^r q_v$  as the set of nodes with rank up to  $r$ . Please note that, by definition,  $Q_{err}^{(|Q_{err}|)} \equiv Q_{err}$ . The overall performance of the set  $Q_{err}$  is computed according to the equation

$$P(Q_{err}) = \frac{1}{N|Q_{err}|} \sum_{r=1}^{|Q_{err}|} O(Q_{err}^{(r)}), \quad (1)$$

where  $O(Q_{err}^{(r)})$  is the average size of the outbreak in the ICM when the set of seeds is given by  $Q_{err}^{(r)}$ . For given  $A_{err}$ , the overall performance is finally obtained by taking the average over all  $V$  realizations of the sets of top spreaders. To account for the stochastic nature of structural noise, we repeat the entire procedure on  $G = 10$  instances of  $A_{err}$  and quantify the spreading performance of top spreaders as the average value over these independent realizations. Please note that the sum on the r.h.s. of eq. (1) allows us to estimate not only the overall performance of the set  $Q_{err}$ , but also the way the set is constructed. The pre-factor appearing on the r.h.s. of eq. (1) is used only to confine values of our performance metric to the interval  $[0, 1]$ . Our focus here is not on measuring the effectiveness of the algorithm used to determine the spreaders, rather on the importance that prior information has in the selection of the top spreaders. Other metrics could be used in place of the one defined in eq. (1). For instance, metrics that consider the identity of the nodes in  $Q_{err}$  with

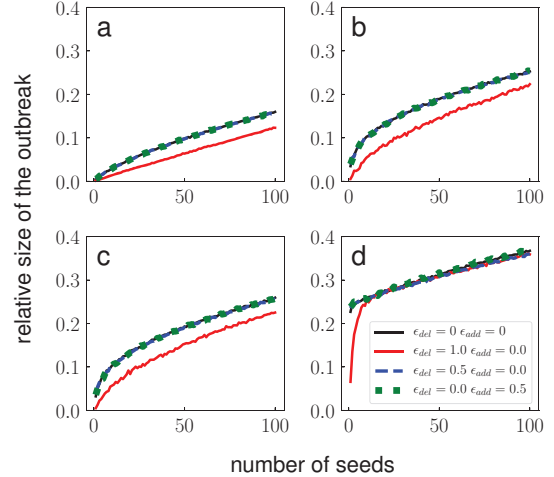


Fig. 1: (Color online) Influence maximization in the presence of structural and dynamical noise. The true network structure analyzed here is given by the e-mail communication network of ref. [23]. (a) Relative size of the outbreak  $O/N$  as a function of the number of seeds found by greedy optimization. The true spreading probability of the ICM is such that  $\phi_{true} = 0.5$ . Prior dynamical knowledge used by the greedy algorithm is not affected by noise, *i.e.*,  $\phi_{err} = \phi_{true} = 0.5$ . The different curves correspond to different levels of noise that affect prior structural information. We consider various combinations of the parameters  $\epsilon_{del}$  and  $\epsilon_{add}$ . (b) Same as in panel (a), but for  $\phi_{true} = 1.0$  and  $\phi_{err} = 0.5$ . (c) Same as in panel (a), but for  $\phi_{true} = \phi_{err} = 1.0$ . (d) Same as in panel (a), but for  $\phi_{true} = 1.5$  and  $\phi_{err} = 1.0$ .

respect to those found in the true set of top spreaders. We believe, however, that this second type of metric may be misleading as the difference in terms of outbreak size between the optimal solution and a slightly less optimal solution may be very small despite a high dissimilarity in terms of the nodes that define the two solutions. In the SM, we actually verified that the identity of the nodes in the set  $Q_{err}$  may be sensitive to the choice of the noise parameters and the specific run of the identification algorithm; instead, the size of the outbreak  $O$  is not much affected by fluctuations.

**Results.** – In fig. 1, we display results obtained for the e-mail communication network originally considered in ref. [23]. In the majority of the cases, the performance of the set of top spreaders appears robust against structural noise. However, as fig. 1 shows, the overall ability to properly select top spreaders may be seriously affected by the level of noise associated with prior information both at the structural and the dynamical levels. Major issues seem to arise when  $\phi_{true} > 1$ , but  $\phi_{err} < 1$  (see fig. 1(d)).

To better characterize the observed trend, we set  $|Q_{err}| = 100$  and quantify the performance defined in eq. (1) for different levels of noise. The results of this analysis are presented in fig. 2. The various panels of the figure refer to different dynamical regimes identified by

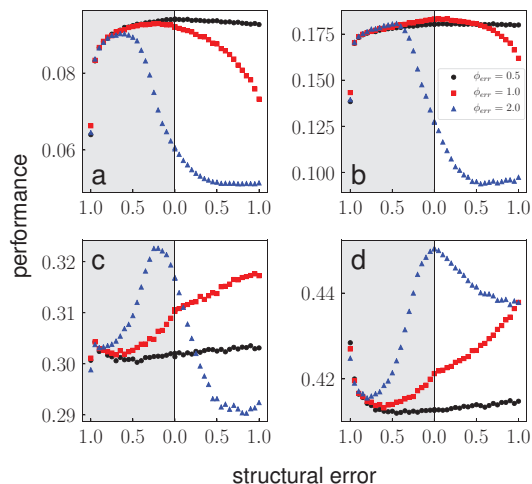


Fig. 2: (Color online) Performance of the top spreaders in the presence of structural and dynamical noise. We consider the same network as in fig. 1. (a) We compute eq. (1) for the set of top spreaders of size  $|Q_{err}| = 100$ , and we plot the value of the performance as a function of the noise level in prior structural information. Performance is measured for  $\phi_{true} = 0.5$ . The shaded part of the plot serves to report results valid for  $0 \leq \epsilon_{del} \leq 1$  and  $\epsilon_{add} = 0$ . The non-shaded part of the graph represents instead results for  $0 \leq \epsilon_{add} \leq 1$  and  $\epsilon_{del} = 0$ . (b) Same as in panel (a), but for  $\phi_{true} = 1.0$ . (c) Same as in panel (a), but for  $\phi_{true} = 1.5$ . (d) Same as in panel (a), but for  $\phi_{true} = 2.0$ .

different values of  $\phi_{true}$ . In every panel, we present three curves, each representing a specific value of  $\phi_{err}$ . Each curve stands for spreading performance  $P(Q_{err})$  as a function of the structural noise parameters  $\epsilon_{del}$  and  $\epsilon_{add}$ . Please consider that, although the two sources of structural noise are never considered active simultaneously, we present them in the same plot for the sake of compactness. The general observed behavior can be summarized as follows. Maximal performance is reached at  $\epsilon_{del} = \epsilon_{add} = 0$ , only for  $\phi_{err} = \phi_{true}$ . The two sources of structural noise affect the choice of the top spreaders differently. Consider first the case  $\epsilon_{del} = 0$ , but  $0 \leq \epsilon_{add} \leq 1$ . Roughly, we see that  $P(Q_{err})$  is a monotonic function of  $\epsilon_{add}$ , decreasing if  $\phi_{err} \geq \phi_{true}$ , and increasing, otherwise. In the region  $\epsilon_{add} = 0$  and  $0 \leq \epsilon_{del} \leq 1$ , instead,  $P(Q_{err})$  is not a monotonic function of the structural error. Further, the trend changes depending on whether the system is in the subcritical or supercritical regime: if  $\phi_{true} \leq 1$ ,  $P(Q_{err})$  is concave; if  $\phi_{true} \geq 1$ ,  $P(Q_{err})$  is convex. In this second regime, it becomes possible to obtain higher performance by adding further structural mistakes. If  $\phi_{err} < \phi_{true}$ , the best performance is achieved for  $\epsilon_{del} = 1$ , essentially using no prior knowledge of the network structure, *i.e.*, seeds are sampled at random.

We analyzed the phenomenon systematically. For different combinations  $\phi_{true}$  and  $\phi_{err}$ , we computed  $\hat{\epsilon}_{del}$  and  $\hat{\epsilon}_{add}$ , *i.e.*, the values of the parameters of structural noise where  $P(Q_{err})$  reaches its maximum. As fig. 3(a) shows, for  $\phi_{err} \simeq \phi_{true}$ ,  $\hat{\epsilon}_{del} \simeq 0$ . However, as soon as

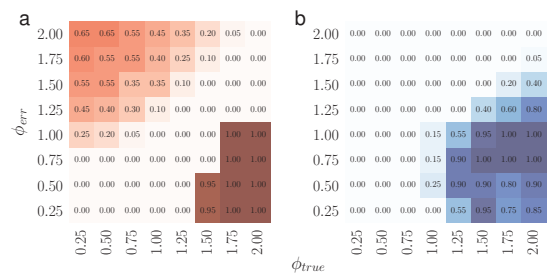


Fig. 3: (Color online) Best values of the structural errors in the presence of dynamical uncertainty. We consider the same network as in fig. 1. (a) We set  $\epsilon_{add} = 0$ , and, for the given dynamical parameters  $\phi_{true}$  and  $\phi_{err}$ , we determine  $\hat{\epsilon}_{del}$ , *i.e.*, the value of the error parameter  $\epsilon_{del}$  that leads to the maximum performance in the prediction of top spreaders. Best estimates of  $\hat{\epsilon}_{del}$  are reported in the cells of the table. The intensity of the background color is proportional to the value of  $\hat{\epsilon}_{del}$ . (b) Same as in panel (a), but for the other source of structural error. We set here  $\epsilon_{del} = 0$  and focus on  $\hat{\epsilon}_{add}$ , *i.e.*, the value of the error parameter  $\epsilon_{add}$  that allows to identify the best performing set of top spreaders.

the error on the dynamical parameter increases, this error may be compensated by making further mistakes in the structure. When structural noise is allowed through the addition of edges instead, noise is helpful only in the supercritical regime (fig. 3(b)).

So far we reported results only for a specific network. However, our main findings are not very sensitive to this choice, in the sense that our qualitative results are very similar for all real networks we analyzed (see SM). The only major difference arises for networks characterized by strong spatial embedding (hence, a specific modular structure identified by very loose intermodule connections [24]). There, the strong asymmetry between the operations of altering the network structure by adding or removing random edges is apparent. This fact is, for instance, visible in fig. 4, where we report the analogue of fig. 2 for the power grid network originally considered in ref. [22]. From the figure, we see that  $P(Q_{err})$  does not behave smoothly around  $\epsilon_{del} = \epsilon_{add} = 0$ . However, the general findings valid for the two different sources of structural noise are almost identical to those valid for networks with no spatial embedding.

The same qualitative results hold for other values of the size of the set of seed nodes  $|Q_{err}|$ , as long as its value is large enough compared to the size of the network  $N$ . In the SM, we report results valid for  $|Q_{err}| = 10$  for the e-mail network considered here in the main text for which  $N = 1133$ . In such a case, the clear pattern of fig. 2 becomes much noisier. For networks of smaller size, we observe that the pattern is already clear even for  $|Q_{err}| = 10$ .

One may explain our results with the following naive argument. For simplicity, let us consider only the case where noise randomly deletes true edges. In our prior knowledge, every true edge becomes invisible with probability  $\epsilon_{del}$ .

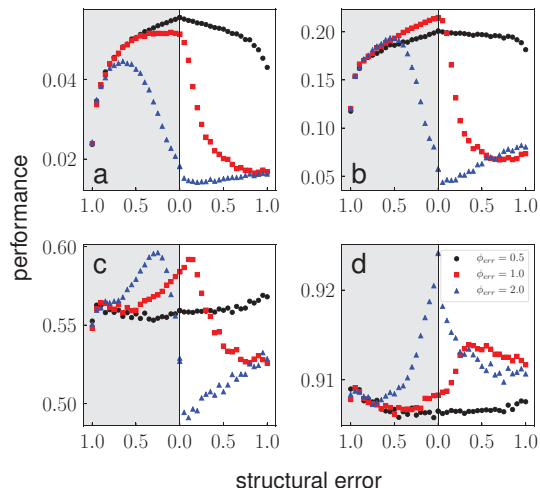


Fig. 4: (Color online) Performance of the top spreaders on a spatially embedded network in the presence of structural and dynamical noise. Same analysis as in fig. 2, but for a different network. Here, the true network structure is given by the US power grid network of ref. [22].

Further, we believe that the ICM has spreading probability  $\phi_{err}$ . In summary, prior knowledge forces us to think that the effective spreading probability on a random edge that we are considering in the true, but unknown, network is  $(1 - \epsilon_{del})\phi_{err}$ , rather than the actual true value  $\phi_{true}$ . Best predictions should be obtained for  $(1 - \epsilon_{del})\phi_{err} \simeq \phi_{true}$ . If  $\phi_{err} > \phi_{true}$ , one can correct the mistake by choosing appropriately  $\epsilon_{del} \in [0, 1]$ . If  $\phi_{err} < \phi_{true}$ , there is no way to satisfy the previous equation. The best performance would be naively expected for  $\epsilon_{del} = 0$ , as this value corresponds to the noise level that minimizes the difference between effective and true spreading probability. However, this is not what we observe in our numerical results, where the best performance is actually achieved for  $\epsilon_{del} = 1$ . The apparent paradox can be solved by accounting for structural correlations. As is well known, true top spreaders in the ICM depend on the critical regime [13]. In the subcritical regime, central nodes are generally better locations for seeds. In the supercritical regime instead, peripheral nodes are selected first. In both regimes, seeds are generally placed on nodes that are not directly connected, as a source of spreading that is too redundant is generally not optimal. If the probability  $\epsilon_{del}$  of random deletion of edges is not very high, then the ranking based on the degree centrality of the nodes is basically unaffected. However, pairs of truly connected nodes appear as disconnected in the noisy version of the network regardless of their degree. As a result, for  $\phi_{err} < 1$  many high-degree nodes are chosen as seeds. However, they may behave poorly in terms of seed set, as they constitute a source of spreading that is too redundant to be optimal. A visual intuition of this structural explanation is provided in fig. 5. There, we consider the true value of the average degree of the set of top spreaders identified

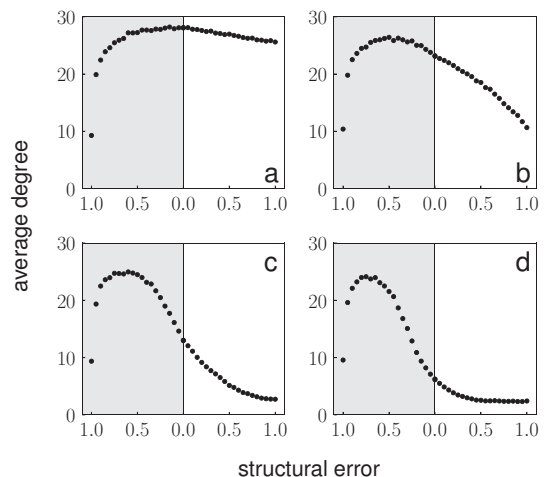


Fig. 5: Average degree of the set of top spreaders. We consider the same network as in fig. 1. (a) Average degree of the set of  $|Q_{err}| = 100$  of top spreaders identified for  $\phi_{err} = 0.5$  and different values of the structural errors  $\epsilon_{del}$  and  $\epsilon_{add}$ . As in fig. 2, we use the left part of the plot, highlighted with a gray-shaded background, to report results valid for  $0 \leq \epsilon_{del} \leq 1$  and  $\epsilon_{add} = 0$ . The non-shaded part of the graph represents instead results for  $0 \leq \epsilon_{add} \leq 1$  and  $\epsilon_{del} = 0$ . (b)–(d) Same as in panel (a), but for  $\phi_{err} = 1.0$ ,  $\phi_{err} = 1.5$ , and  $\phi_{err} = 2.0$ , respectively.

using noisy information. Each panel corresponds to a different value of  $\phi_{err}$ . Please note that the value of the average degree measured at  $\epsilon_{del} = \epsilon_{add} = 0$  is the one that corresponds to the true set of optimal seeds for the dynamical regime  $\phi_{true} = \phi_{err}$ . As expected, for the subcritical regime, the identified set of top spreaders has high average degree and the structural noise does not affect much the value of this variable, except when  $\epsilon_{del} \simeq 1$ . In the supercritical regimes instead, the best performance is achieved for sets with low values of the average degree, comparable with the average degree of the network. Structural noise changes dramatically the set of seeds, especially in the region  $0 < \epsilon_{del} < 1$ . However, in the regime of very large noise, the average degree of the seed set is basically equal to the average degree of the network as nodes are chosen using (almost) no structural information.

**Conclusions.** – In this paper, we considered a simple yet practically relevant scenario. We assumed that prior information used in the solution of the influence maximization problem is affected by some noise, and we studied how the quality of the solution found using noisy information deteriorates as a function of the noise intensity. Our main finding is that the quality of the solution always decreases monotonically with noise, if structural and dynamical noise are considered independently. However, when both sources of noise act simultaneously, one of them can compensate the disruptive effect of the other. In essence, noise affecting dynamical information may be suppressed by additional noise at the structural level, or vice versa.

This fact is particularly apparent when structural noise is such that random edges of the original network disappear with a certain probability. As this is a plausible model of error that may affect our knowledge of the true network structure [14], our results may be important in real-world applications. More in general, the approach presented here may be used to understand how incomplete and/or erroneous information at the level of network structure and dynamics affects our ability to solve optimization problems in a meaningful way.

\* \* \*

The authors thank C. CASTELLANO for critical reading of the manuscript. SE and FR acknowledge support from the National Science Foundation (CMMI-1552487). AF and FR acknowledge support from the US Army Research Office (W911NF-16-1-0104). AF acknowledges support from the Science Foundation Ireland (16/IA/4470).

## REFERENCES

- [1] RATKIEWICZ J., CONOVER M., MEISS M., GONÇALVES B., PATIL S., FLAMMINI A. and MENCZER F., *Truthy: Mapping the spread of astroturf in microblog streams*, in *Proceedings of the 20th International Conference Companion on World Wide Web* (ACM) 2011, pp. 249–252.
- [2] DOMINGOS P. and RICHARDSON M., *Mining the network value of customers*, in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '01* (ACM, New York, NY, USA) 2001, pp. 57–66.
- [3] KEMPE D., KLEINBERG J. and TARDOS E., *Maximizing the spread of influence through a social network*, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '03* (ACM, New York, NY, USA) 2003, pp. 137–146.
- [4] CHEN W., WANG Y. and YANG S., *Efficient influence maximization in social networks*, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '09* (ACM, New York, NY, USA) 2009, pp. 199–208.
- [5] KITSACK M., GALLOS L. K., HAVLIN S., LILJEROS F., MUCHNIK L. STANLEY H. E. and MAKSE H. A., *Nat. Phys.*, **6** (2010) 888.
- [6] BAUER F. and LIZIER J. T., *EPL*, **99** (2012) 68007.
- [7] KLEMM K., SERRANO M. Á., EGUÍLUZ V. M. and SAN MIGUEL M., *Sci. Rep.*, **2** (2012) 292.
- [8] CHEN D., L. L., SHANG M.-S., ZHANG Y.-C. and ZHOU T., *Physica A: Stat. Mech. Appl.*, **391** (2012) 1777.
- [9] CHEN D.-B., XIAO R., ZENG A. and ZHANG Y.-C., *EPL*, **104** (2013) 68006.
- [10] DE ARRUDA G. F., BARBIERI A. L., RODRÍGUEZ P. M., RODRIGUES F. A., MORENO Y. and COSTA L. D. F., *Phys. Rev. E*, **90** (2014) 032812.
- [11] MORONE F. and MAKSE H. A., *Nature*, **524** (2015) 65.
- [12] RADICCHI F. and CASTELLANO C., *Phys. Rev. Lett.*, **120** (2018) 198301.
- [13] RADICCHI F. and CASTELLANO C., *Phys. Rev. E*, **95** (2017) 012318.
- [14] NEWMAN M. E. J., *Nat. Phys.*, **14** (2018) 542.
- [15] SAITO K., NAKANO R. and KIMURA M., *Prediction of information diffusion probabilities for independent cascade model*, in *Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (Springer) 2008, pp. 67–75.
- [16] DALL’ASTA L., ALVAREZ-HAMELIN I., BARRAT A., VÁZQUEZ A. and VESPIGNANI A., *Phys. Rev. E*, **71** (2005) 036135.
- [17] LESKOVEC J. and FALOUTSOS C., *Sampling from large graphs*, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '06* (ACM, New York, NY, USA) 2006, pp. 631–636.
- [18] PASTOR-SATORRAS R., CASTELLANO C., VAN MIEGHEM P. and VESPIGNANI A., *Rev. Mod. Phys.*, **87** (2015) 925.
- [19] NEMHAUSER G. L., WOLSEY L. A. and FISHER M. L., *Math. Program.*, **14** (1978) 265.
- [20] GRASSBERGER P., *Math. Biosci.*, **63** (1983) 157.
- [21] RADICCHI F. and CASTELLANO C., *Phys. Rev. E*, **93** (2016) 062314.
- [22] WATTS D. J. and STROGATZ S. H., *Nature*, **393** (1998) 440.
- [23] GUIMERÀ R., DANON L., DÍAZ-GUILERA A., GIRALT F. and ARENAS A., *Phys. Rev. E*, **68** (2003) 065103.
- [24] FAQEEH A., MELNIK S., COLOMER-DE SIMÓN P. and GLEESON J. P., *Phys. Rev. E*, **93** (2016) 062308.