I. INTRODUCTION

Every day, we witness the dissemination of new pieces of information in social networks [1][5]. Few of them become widespread; the vast majority, however, diffuse only over a vanishing portion of the network. Are there a priori identifiable features that allow for the early prediction of the outcome of a spreading process in a network? Many studies have pointed out that the “quality” or “attractiveness” of the information might have an effect on how far it may spread [1][6]. In mathematical models of information spreading, the notion of quality is typically quantified in terms of the probability of spreading events along individual edges in the social network. However, the spreading probability of individual edges is not the only key factor that determines the fate of a piece of information spreading in a network. The nodes that act as seeds for the spreading process may play a role that is more important than the actual probability to spread information along social contacts. Intuitively, if the diffusion process is seeded by central nodes, then the piece of information may reach large popularity; on the other hand, a piece of information originated from peripheral nodes is much less likely to become widespread.

The problem of selecting the best set of seed nodes for a spreading process in a network has been traditionally named as the problem of influence maximization. The problem is generally considered under the strong assumption of having full and exact knowledge of both the network topology and the spreading dynamics. We will adopt this line here too, although we remark that such an assumption is at least optimistic and may potentially lead to significant mistakes in the identification of the true influential spreaders [7]. The function that is optimized in influence maximization is the average value of the outbreak size. The optimization problem is solved for a given size of the seed set, generally much smaller than the network size. The problem was first formulated by Domingos and Richardson [8], and later generalized by Kempe et al. [9]. In particular, Kempe et al. showed that influence maximization is a NP-hard problem, exactly solvable for very small networks only. Also, Kempe et al. demonstrated that for specific models of opinion spreading, such as the independent cascade and the linear threshold models, the average outbreak size is a submodular function, and thus greedy optimization algorithms allow to find, in polynomial time, approximate solutions that are less than a factor \( (1 - 1/e) \) away from the true optimum [10].

The greedy algorithm actively uses information about the topology of the network and the dynamical rules of the spreading model. After the seminal work by Kempe et al., other similar greedy techniques for approximating solutions to the influence maximization problem have been proposed [11][14]. As all these algorithms require knowledge of the model at the basis of the spreading process, often obtained through numerical simulations, they all suffer from the limitation of being applicable to small-medium sized networks only. We remark that some attempts of greedy-like algorithms applicable to large networks have been made [15][16]. Those attempts, however, rely on approximate estimations of the outcome of numerical simulations, thus leading to solutions to the influence maximization problem that are generally inferior to the solutions obtained with straight greedy optimization.

On large networks, like those of interest in practical applications, solutions to the influence maximization problem are generally obtained via heuristic methods. The literature is full of examples [17][23]. Heuristic methods use complete
information about the network structure, but they completely neglect information about the dynamical model of spreading. They are generally much faster than greedy algorithms, but clearly less effective. Their main limitations are two-fold. On the one hand, heuristic methods are characterized by the inability to account for the combined effect that seeds may have in a complex spreading process, as the set of influential nodes is built combining the best individual spreaders and their influence sets may be strongly overlapping. On the other hand, being based on purely topological properties, heuristic methods lack sensitivity to the features of the spreading dynamics and the variation of the associated parameters. Given the wealth of heuristic methods that have been proposed to identify influential nodes in networks, how different these methods are in terms of performance? Even more important, how far is the performance of the best heuristic methods from optimality, at least the achievable optimality provided by greedy algorithms? We realized that no clear answer to these fundamental questions can be found in current literature, and we decided to fill this gap of knowledge here.

The present paper reports on a systematic test of 16 heuristic methods that have been proposed to approximate solutions to the influence maximization problem. Our analysis is based on a corpus of 100 real-world networks, and performance of the various heuristic methods is quantified for SIR-like spreading processes. Despite the various methods rely on rather different centrality metrics, we find that many of them all are able to achieve comparable performances. When used to select the top 5% initial seeds of spreading in real-networks, the best performing methods show levels of performance that are within 90% from those achievable by greedy optimization, so that the room for potential improvement appears small. We show that one way to achieve better performances is relying on hybrid methods that combine two or more centrality metrics together. We validate this final result on a small set of large-scale networks.

II. METHODS

A. Networks

In this study, we focus most of our attention on a corpus of 100, undirected and unweighted, real-world networks. Sizes of these networks range from 100 to 30,000 nodes, and their density varies between 0.0001 and 0.25. The corpus is composed of networks of small to medium size on purpose, as these allow for the application of greedy optimization in the solution of the influence maximization problem. We consider networks from different domains. Specifically, our corpus of networks include 61 social, 16 technological, 12 information, 8 biological, and 3 transportation networks. Details about the analyzed networks can be found in the SM1 [24]. In the final part of the paper, we validate some of our findings on 9 large real-world social and information networks with sizes ranging from 50,000 to slightly more than 1,000,000. Details are provided in Table III.

B. Spreading dynamics

We concentrate our attention on the Independent Cascade Model (ICM) [9]. This is a very popular model in studies focusing on the influence maximization problem. The ICM is a simplified version of the Susceptible-Infected-Recovered (SIR) model [25]. Nodes can be in either of the three states S, I, or R. At the beginning of the dynamics, all nodes start in the S state except for those who are selected to be the initial spreaders, which are assigned to the I state. At each step of the model, all nodes in state I try to infect their neighbors in state S with probability $p$; then, they recover immediately, by changing their states from I to R. Nodes in state R never change their state and no longer participate to the spreading dynamics. The dynamics continues until there are no nodes left in state I. The size of the outbreak is calculated by counting the number of nodes that ended up in state R at the end of the spreading dynamics. As the spreading from one node to another happens with probability $p$, the model has a stochastic nature. To properly account for the stochastic nature of the model, all our results are obtained as average values over 50 independent numerical simulations for every given initial condition.

C. Methods for the selection of influential spreaders

In total, we consider 18 methods for the identification of influential spreaders in networks (see Table I). Each method outputs a list of nodes in a specific order from the most influential to the least influential node. We use this rank to construct, in a sequential manner, the set of the top spreaders according to a particular method. The various methods take as input different type/amount of information, and make use of rather different types of rankings. As a consequence, the computational complexity of the various methods may be significantly different. For illustrative purposes, we decided to group the 18 methods for the selection of influential spreaders into four main groups.

The group of baseline methods is formed by the methods greedy and random. The greedy algorithm is the best performing method available on the market, thus providing an upper bound for the performance of all other methods. The greedy algorithm uses all available information about network topology and spreading dynamics. For instance, the algorithm provides different solutions depending on the value of the spreading probability $p$. For the greedy method applied to the ICM, we rely on the Chen et al.'s [12] algorithm, which makes use of the mapping between ICM and bond percolation to obtain faster results regarding the simulations of the spreading process. The random method instead represents a lower bound for the performances of other methods. The method just outputs nodes of the network in random order, de facto neglecting any prior information regarding system topology and dynamics.

The remaining 16 of the 18 methods are purely topological methods in the sense that they rely on heuristics that are calculated using full knowledge of the network structure, but no information at all about spreading dynamics. According to
Table I. Methods for the selection of influential spreaders. We list basic details of all the methods for the detection of influential spreaders in complex networks that we consider in this study. Each row of the table refers to a specific method. From left to right, we report the full name of the method, the abbreviation of the method name, the reference of the paper where the method was introduced, and the computational complexity of the method. Computational complexities reported in the table are obtained under the realistic assumption that methods are applied to sparse networks where the number of edges scales linearly with the network size. Methods are further grouped into different categories, i.e., baseline, local, global and intermediate, depending on their properties.

| Group      | Method                      | Abbrev. | Ref | Complexity |
|------------|-----------------------------|---------|-----|------------|
| Baseline   | Greedy                      | G       | [12] | cubic      |
|            | Random                      | R       | -   | constant   |
| Local      | Degree                      | D       | -   | linear     |
|            | Adaptive Degree             | AD      | [12] | linear     |
| Global     | Betweenness                 | B       | [26] | quadratic  |
|            | Closeness                   | C       | [27] | quadratic  |
|            | Eigenvector                 | E       | [28] | linear     |
|            | Katz                        | K       | [29] | linear     |
|            | PageRank                    | PR      | [30] | linear     |
|            | Non-backtracking            | NB      | [31] | linear     |
|            | Adaptive NB                 | ANB     | [32] | quadratic  |
| Intermediate | k-shell                      | KS      | [33] | linear     |
|            | LocalRank                   | LR      | [34] | linear     |
|            | CoreHD                      | CD      | [36] | linear     |
|            | Collective Influence, $\ell = 1$ | CII | [37] | linear     |
|            | Collective Influence, $\ell = 2$ | CII | [37] | linear     |
|            | Expl. Immunization          | EI      | [38] | linear     |

For every set of top $m$ nodes, we indicate as $S^m_t$ the set of top $m$ nodes identified by method $m$ in instance $t$ of the method and for a given network with $N$ nodes. For every set $S^m_t$, we run 50 different times the ICM model, and measure the average value of the outbreak

Figure 1. Relative size of the outbreak as a function of the relative size of seed set for the email communication network of Ref. [42]. To obtain relative values, we divide outbreak size and seed set size by the total number of nodes in the network. Relative measures allow for an immediate comparison across networks with different sizes. We compare the performance of different methods for the selection of influential nodes. Outbreak size is calculated for ICM dynamics at critical threshold $p_c = 0.056$. To avoid confusion, we display results only for a subset of the methods considered in the paper.

Some of the selection methods described above are stochastic in the sense that they may generate a different ranking for the nodes at each run. To account for this fact, we apply every stochastic method $R = 10$ independent times to generate rankings for the nodes. We consider each of these rankings to sequentially construct sets of top spreaders. Specifically, we indicate as $S^{(r)}_m$ the set of top $\ell N$ spreaders identified by method $m$ and for a given network with $N$ nodes. For every set $S^{(r)}_m$, we run 50 different times the ICM model, and measure the average value of the outbreak
size $O(S_m^{(T)})$. We then repeat the operation for every instance $r$ of the method, and take the average over the $R$ potentially different sets, namely

$$V_m^{(i)} = \frac{1}{R} \sum_{r=1}^{R} O(S_m^{(i)}) .$$

Fig. 1 displays how the relative size of the outbreak $V_m^{(i)}/N$ grows as function of the relative seed set size $t$ for some of the methods for the identification of top spreaders considered in this paper. The figure clearly shows that the greedy and random algorithms are good baselines for the performances of the other methods. For instance, the greedy algorithm outperforms all other methods. This result is confirmed across the entire corpus of networks we analyzed in this paper (see SM1 [24] and SM2 [41]). In a few networks, some heuristic methods are able to slightly outperform the greedy algorithm, but the difference is so little that it can be attributed to statistical fluctuations. Similarly, all methods perform better than the random selection method, although there are quite a few cases where randomly selecting seeds perform as well as selecting seeds according to some topological heuristic.

As a measure for the performance of method $m$ in the identification of the top $T N$ influential spreaders of a given network, we evaluate the area under the curves of Fig. 1 up to a pre-imposed $T$ value

$$q_m^{(T)} = \frac{1}{N} \int_0^T dt V_m^{(i)} .$$

(1)

As the size of set of top spreaders is linearly dependent on the size of the network $N$, we can easily aggregate results obtained over the entire corpus of real-world networks at our disposal. Specifically, results in the main paper are obtained for $T = 0.05$. We report results for $T = 0.1$ in the SM2 [41]. No significant differences between the two cases are apparent. As some of the methods considered in the paper are characterized by large computational complexity (see Table I), we couldn’t consider $T > 0.1$. We note, however, that studying the performance of methods for the identification of influential spreaders has a meaning only for small $T$ values, given that in practical applications the seeding is generally performed on a vanishing portion of the system. Also, we test the validity of all results using $V_m^{(T)}$ as a main metric of performance, instead of its integral of Eq. (1). Results are reported in the SM2 [41]. No significant changes with respect to the results presented here in the main paper are apparent.

As the greedy algorithm provides an upper bound for the performance of the other methods, we use it as a term of comparison for all other methods in our systematic analysis. We consider two main metrics of performance. The first measure is based on a comparison between the outbreak size obtainable by a method compared to the one obtained using the greedy identification method. Specifically, given a network, we first compute

$$g_m^{(T)} = \frac{q_m^{(T)}}{q_G^{(T)}} ,$$

(2)

where we used the abbreviation $q_m^{(T)}$ to indicate the expression of Eq. (1) for the greedy algorithm, i.e., $m = G$. Then, we evaluate the performance relative to greedy for all networks in our dataset, and summarize the results in Fig. 2 where we display the cumulative distribution of this quantity for some of the methods. To obtain a single number for the performance of the method over the entire corpus of networks, we define the overall performance $(g_m^{(T)})$ given by the average value of the metric defined in Eq. (2) over all real networks in the dataset.

The second metric of performance instead neglects the size of the outbreak, and focuses only on the identity of the nodes identified by the method $m$. For the actual solution of the problem of influence maximization, this second metric is clearly much less important than the one previously considered. However, the metric can tell us something more about the topological properties of the set of top spreaders in networks. Given a network, we evaluate the frequency $f_m^{(T)} = f_m^{(T)}$ of every node $i$ to be in the set of top $T N$ spreaders according to method $m$ over $R = 10$ runs of the algorithm. We then compute the precision of the method relative to the greedy algorithm as

$$r_m^{(T)} = \frac{1}{T N} \sum_{i=1}^{N} f_m^{(T)} f_G^{(T)} .$$

(3)

We note that Eq. (3) can be used to measure the self-consistency of the greedy method by setting $m = G$. The cumulative distribution of the precision metric defined in Eq. (3) across the entire network dataset is displayed in Fig. 3. The plot shows high level of precision between some methods and
the greedy algorithm. The random selection method generates a distribution well peaked around the value $T$. We characterize the generic method $m$ with a metric of overall precision $\langle r_m \rangle$ as the average value of the precision defined in Eq. [3] over the entire corpus of real networks. The value of this metric tells us how much the method $m$ is similar to the baseline provided by the greedy algorithm in the identification of the top spreaders across the entire corpus of networks at our disposal.

III. RESULTS

A. Individual methods

Armed with the metrics defined in the section above, we test the various methods for the identification of influential spreaders for ICM dynamics over the entire corpus of real networks at our disposal. We remark that both the identity and performance of the true set of influential spreaders may be dependent on the actual value of the spreading probability $p$ in the ICM model, so that the performance of the various seed selection methods needs to be evaluated at different values of the spreading probability $p$. For instance, for the extreme cases $p = 0$ and $p = 1$, predictions are trivial in the sense that all methods have exactly the same performance in terms of outbreak size. The prediction of methods performance is instead non trivial when the uncertainty of the spreading outcome is maximal. For this reason, we focus our attention on ICM dynamics around the critical threshold $p = p_c$. To perform the analysis, we first evaluate the critical threshold values $p_c$ for every network in the database. Specifically, we rely on mapping between bond percolation and the ICM, and we apply the Newman-Ziff algorithm to evaluate $p_c$. $p_c$ values for the various networks are reported in the SM1 [24]. We then consider ICM dynamics for three distinct values of $p$: (i) sub-critical regime at $p = p_c/2$; (ii) critical regime at $p = p_c$; (iii) supercritical regime at $p = 2p_c$.

Results of our analysis are summarized in Fig. 4. Every method is used to identify the set of top $N$ nodes in the networks, with $T = 0.05$. In the figure, we represent results for each method $m$ in the plane $(\langle g_m \rangle, \langle r_m \rangle)$. Please note that we dropped the suffix $T$ to simplify the notation. We remark that the performance of every method $m$ is measured in relation to the performance of the greedy method, i.e., $m = G$. By definition, we have $\langle g_G \rangle = 1$; we find instead that the self-consistency score is $\langle r_G \rangle < 1$ meaning that optimal sets identified by the greedy algorithm have some degree of variability. An interesting finding is the absence of a strong dependence of $\langle r_G \rangle$ from the dynamical regimes of the ICM. The other important reference point in the plane is given by the random method ($m = R$). By definition, we have that $\langle r_R \rangle = T = 0.05$. $\langle g_R \rangle$ values instead strongly depend on the dynamical regime.

In the subcritical regime (see Fig. 4a), the two metrics $\langle g_m \rangle$ and $\langle r_m \rangle$ are tightly related one to the other. Adaptive degree ($m = AD$) outperforms all other methods in both metrics. Other methods that perform very well are those based on algorithms relying on the Degree ($m = D$), Adaptive Non-Backtracking ($m = ANB$) and PageRank ($m = PR$) centralities, as well as those based on the CoreHD ($m = CD$) and Collective Influence ($m = CI$) algorithms. Similar considerations apply to the critical regime (Fig. 4b). The most significant change with respect to the subcritical regime is a slight decrease of range of values for the performance metric of the algorithms. In the supercritical regime (Fig. 4c), there is no longer a proper distinction between the various methods in terms of performance.

A remarkable feature emerging from Fig. 4 is that the overall performance is rather high. For most of the methods values are above 0.9 for all values of $p$, and even random selection provides a performance always larger than 0.6. This observation somehow helps to properly weigh the importance of greedy algorithms for influence maximization: why their solutions are guaranteed to be not too far from the true optimum, their performance can be almost achieved by simple and much more easily implemented purely topological methods.

The similarity in the performance between the various methods can be deduced by a straight pair-wise comparison between the sets of top influential nodes identified by the various methods across the entire corpus of real networks at our disposal. The results of this analysis are summarized in Fig. 5. Top-performing methods provide sets of influential nodes very similar to each other; methods with low performance instead generally identify influential nodes that are rarely selected by any other method.

In the SM2 [41], we repeat the same exercise by computing the performance scores restricted to different subsets of the whole corpus of networks. The subsets correspond to
Barabasi-Albert (BA) model [45]. Results are very similar to the main outcome of the analysis. We further consider artificial networks created with the Barabasi-Albert (BA) model [45]. Results are very similar to those obtained on real-world networks. In summary, it seems that the main results of the paper are unchanged by the nature/type of the network substrate where spreading is occurring.

B. Hybrid methods

In this section, we report on the performance of hybrid methods for the identification of top spreaders in the network obtained from linear combinations of the individual methods considered so far. Specifically, we first select a certain number of individual methods to form an hybrid method \( \mathcal{H} = \{ m_1, m_2, \ldots, m_m \} \). We associate to every node \( i \) in a given network a score \( s^{(\mathcal{H})}_m \) that is a linear combination of the scores associated with individual methods, namely

\[
s^{(\mathcal{H})}_m = \sum_{m \in \mathcal{H}} c_m s^{(m)}_m .
\]

In Eq. (4), \( s^{(m)}_m \) is the normalized score of node \( i \) in the network according to the topological metric used by method \( m \). The normalization \((L^2\text{-norm})\) has the purpose of making scores of comparable magnitude across methods. The best estimates of the linear coefficients \( c_m \) are then obtained using information from the greedy algorithm. We use linear regression to find the best linear fit between \( s^{(\mathcal{H})}_m \) and \( s^{(G)}_m \), i.e., the probability that node \( i \) is identified by the greedy algorithm in the set of top \( T N \) influential nodes in the network. Best estimates of the coefficients are obtained relying on a training set composed of 80% of networks randomly chosen out of the corpus of real networks at our disposal. We then test the hybrid method \( \mathcal{H} \) on the remaining 20% of the corpus, where we measure overall performance and overall precision. We replicate the entire networks from the same domain (e.g., social, technological, transportation); we do not find any significant change in the main outcome of the analysis.

Figure 4. Performance and precision of methods for the identification of influential spreaders in real networks. Results are based on the systematic analysis of 100 real-world networks. For each network, we first evaluate the critical value of the spreading probability \( p_c \) for ICM dynamics. Then, we consider the analysis for three distinct phases of spreading: (a) \( p = p_c/2 \), (b) \( p = p_c \), (c) \( p = 2p_c \). Each point in the various panels corresponds to one method. Every method is used to identify the top \( T N \), with \( T = 0.05 \), spreaders in the networks. For clarity of the figure, methods are identified by the same abbreviations as those defined in Table I. Methods are characterized by the metrics of performance defined in the paper. Both these metrics relate the performance of a generic method \( m \) to the one of the greedy algorithm. Overall performance \( \langle r_m \rangle \) is a metric of performance that relies on the size of the outbreak associated with the set of influential spreaders identified by the method compared to the typical outbreak obtained with the greedy algorithm. Overall precision \( \langle p_m \rangle \) instead quantifies the overlap between the sets of spreaders identified by a method and those identified by the greedy algorithm.

Figure 5. Pairwise comparison among methods for the identification of influential spreaders. For every pair of methods \( m_1 \) and \( m_2 \), we evaluated the overlap \( r^{(T)}_{m_1,m_2} \) among the two sets of top \( T N \) influential spreaders found by the methods in the network using a precision metric similar to the one of Eq. (3), i.e.,

\[
r^{(T)}_{m_1,m_2} = \frac{1}{T N} \sum_{i=1}^{T N} \sum_{j=1}^{T N} \delta(m_1, i) \delta(m_2, j).
\]

We then estimated the average value of the precision over the entire corpus of real networks at our disposal. In the figure, dark colors correspond to high values of precision; low precision values are represented with light colors. Acronyms of the methods are defined in Table I. Methods are listed in the table according to the same order as they appear in Table I.
We consider several hybrid methods consisting in the combination of two and three individual centrality metrics. In general, we combine together centrality methods that differ on the basis of their classification in local, global and intermediate regimes (see Table II). Results for some hybrid methods are reported in Table II. Several remarks are in order. First, with respect to the case of individual methods, there is an increase in the measured values of the overall precision $\langle r_m \rangle$. This tells us that the coefficients learned from the training set can be meaningfully used on other networks to mimic greedy optimization in terms of topological features only. The overall performance $\langle g_m \rangle$ of hybrid methods increases too; improvements beat even by $2 - 5\%$ the best individual methods. Second, when similar individual methods are combined together into an hybrid method, one of the two gets the biggest part of the weight compared to the other. For example, the hybrid method $H = \{AD, B\}$ learned from data is almost a pure AD method in both the subcritical and critical regimes. Third, the coefficients of the linear combination of Eq. $4$ can also be negative. For example, for the hybrid method $H = \{AD, PR, LR\}$ in the critical regime, $c_{LR} < 0$. Thanks to this fact, the method outperforms in both the critical and subcritical regimes all other methods considered in this paper.

To validate the use of hybrid methods for the identification of influential spreaders, we apply the top-performing hybrid method $H = \{AD, PR, LR\}$ to large social and information networks. Results are reported in Table III. These networks are too big for the application of greedy optimization, thus the performance of the hybrid method is compared to the one of the method AD by taking the ratio $\langle g_H \rangle / \langle g_{AD} \rangle$. Please note that AD is one of the best individual methods for the identification of influential spreaders according to our analysis on the corpus of small/medium networks. When applying the hybrid method to large networks, we use the same values of the linear coefficients learned from small/medium networks and listed in Table II. Overall, we see that the hybrid method generates improvements in the detection of influential spreaders compared to the simple AD method. Improvements are almost negligible in the subcritical regime. They are instead significant in both the critical and supercritical dynamical regimes, although in the latter case there are wide variations, with striking performance decrease for some networks. On average, we register improvements of $2 - 5\%$. These values are in line to those that can be measured in the corpus of small/medium networks, thus providing additional support to the robustness and generality of our finding. It should be stressed that the hybrid method uses a slightly larger amount of information than the one at disposal of the individual AD method. This might be at the root of the observed performance increase. As a matter of fact, linear coefficients change their value depending on the dynamical regime, so the ranking of the nodes. On the other hand, the improvement in effectiveness doesn’t cause drawbacks in efficiency. Linear coefficients of the various dynamical regimes are given. Also, the computational complexity of estimating numerically the critical threshold $p_c$ scales linearly with system size. De facto, the computational complexity of the overall hybrid method is the same as the one of the indi-
The goal of this paper was to comparatively analyze the performances of heuristic methods aimed at the identification of influential spreaders in networks. We focused our attention on the spreading dynamics modeled by the independent cascade model, and studied a total of 16 methods for the identification of influential spreaders that are being used widely in influence maximization studies. We performed a systematic comparison between the various methods by means of extensive numerical experiments on a large corpus of 100 real-world networks. We further drew upper- and lower-bounds for the performance values achievable in the problem by using respectively results from greedy optimization and random selection. We found that the performance of many simple heuristic methods is not far from that of the more computationally costly greedy algorithm. In this framework, the simplest and most effective strategy among those already on the market that can be used to identify top spreaders in large networks is the adaptive degree centrality. The method based on adaptive degree centrality displays an overall performance score that is 96% of the upper-baseline value in the critical regime of spreading, if used to select a set of top spreaders with size equal to 5% of the entire network. Several other methods have comparable performances to adaptive degree centrality. The overlap between influential spreaders selected by heuristic methods and by the greedy algorithm is considerably lower, but this is not surprising given the NP-complete nature of the optimization problem. We finally found that a potential way to get closer to optimality consists in combining different centrality metric to create hybrid methods. We found that some combinations of three metrics are able to achieve 98% of the upper-baseline value in the critical regime of spreading.

Table III. Identification of influential spreaders in large networks. We compare the performance of the hybrid method $H = \{AD, PR, LR\}$ with the individual method AD. For the hybrid method, we use the values of the coefficients reported in Table II. From left to right, we report the name of the network, number of nodes in the giant component, number of edges in the giant component, critical value $p_c$ of the spreading probability, references to studies where the network was first analyzed, url where network data were downloaded, value of the ratio $\langle g_H \rangle/\langle g_{AD} \rangle$ between the performance metric of the hybrid method $H = \{AD, PR, LR\}$ and the one of the individual method AD for the subcritical, critical and supercritical regimes. The bottom two lines in the table report, for each dynamical regime, average values of the ratios $\langle g_H \rangle/\langle g_{AD} \rangle$ over the set of large networks and over the corpus of 100 networks considered in the rest of the paper.

| Network         | $N$  | $E$      | $p_c$ | Ref. | url | $\langle g_H \rangle/\langle g_{AD} \rangle$ |
|-----------------|------|----------|-------|------|-----|---------------------------------------------|
|                 |      |          |       |      |     | Subcrit. Critical Supercrit.                |
| Slashdot        | 51,083 | 116,573  | 0.0262 | [46] | [47] | 1.003 1.017 1.062                           |
| Gnutella, Aug. 31, 2002 | 62,561 | 147,878  | 0.0956 | [48] | [49] | 1.009 1.040 1.039                           |
| Epinions        | 75,877 | 405,739  | 0.0062 | [47] | [50] | 1.012 1.057 1.130                           |
| Flickr          | 105,722 | 2,316,668 | 0.0142 | [47] | [51] | 1.007 1.082 1.242                           |
| Gowalla         | 196,591 | 950,327  | 0.0073 | [47] | [52] | 1.011 1.024 1.066                           |
| EU email        | 224,832 | 339,925  | 0.0119 | [47] | [53] | 1.002 1.009 0.923                           |
| Web Stanford    | 255,265 | 1,941,926 | 0.0598 | [47] | [54] | 1.009 1.031 1.035                           |
| Amazon, Mar. 2, 2003 | 262,111 | 899,792  | 0.0940 | [47] | [55] | 1.008 1.025 0.994                           |
| YouTube friend. net.| 1,134,890 | 2,987,624 | 0.0063 | [47] | [56] | 1.004 1.013 0.952                           |

Average on large networks

| $\langle g_H \rangle/\langle g_{AD} \rangle$ |
|---------------------------------------------|
| 1.007 1.033 1.049 |

Average on the corpus of 100 networks

| $\langle g_H \rangle/\langle g_{AD} \rangle$ |
|---------------------------------------------|
| 0.998 1.020 1.043 |

IV. CONCLUSIONS

ACKNOWLEDGMENTS

$\S$E and FR acknowledge support from the National Science Foundation (CMMI-1552487). FR acknowledges support from the US Army Research Office (W911NF-16-1-0104).

[1] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, “Truthy: mapping the spread of astroturf in microblog streams,” in Proceedings of the 20th international conference companion on World wide web, pp. 249–252, ACM, 2011.

[2] D. Acemoglu, A. Ozdaglar, and A. ParandehGheibi, “Spread of (mis) information in social networks,” Games and Economic Behavior, vol. 70, no. 2, pp. 194–227, 2010.

[3] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, “The spreading of misinformation online,” Proceedings of the National Academy of Sciences, vol. 113, no. 3, pp. 554–559, 2016.

[4] D. Centola, “The spread of behavior in an online social network experiment,” science, vol. 329, no. 5996, pp. 1194–1197, 2010.
[5] K. Lerman and R. Ghosh, “Information contagion: An empirical study of the spread of news on digg and twitter social networks,” *Icwsm*, vol. 10, pp. 90–97, 2010.

[6] D. Notarmuzi and C. Castillo, “Analytical study of quality-biased competition dynamics for memes in social media,” *EPL (Europhysics Letters)*, vol. 122, no. 2, p. 28002, 2018.

[7] Š. Erkol, A. Faqeeh, and F. Radicchi, “Influence maximization in noisy networks,” *EPL (Europhysics Letters)*, vol. 123, no. 5, p. 58007, 2018.

[8] P. Domingos and M. Richardson, “Mining the network value of customers,” in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 57–66, ACM, 2001.

[9] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, ACM, 2003.

[10] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, “An analysis of approximations for maximizing submodular set functions—i,” *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[11] J. Leskovec, A. Krause, C. Guestrin, J. VanBriesen, and N. Glance, “Cost-effective outbreak detection in networks,” in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 420–429, ACM, 2007.

[12] W. Chen, Y. Wang, and S. Yang, “Efficient influence maximization in social networks,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 199–208, ACM, 2009.

[13] A. Goyal, W. Lu, and L. V. Lakshmanan, “Celf++: optimizing the greedy algorithm for influence maximization in social networks,” in *Proceedings of the 20th international conference companion on World wide web*, pp. 47–48, ACM, 2011.

[14] S. Cheng, H. Shen, J. Huang, G. Zhang, and X. Cheng, “Statigreedy: solving the scalability-accuracy dilemma in influence maximization,” in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 509–518, ACM, 2013.

[15] H. T. Nguyen, M. T. Thai, and T. N. Dinh, “Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks,” in *Proceedings of the 2016 International Conference on Management of Data*, pp. 695–710, ACM, 2016.

[16] Y. Hu, S. Ji, Y. Jin, L. Feng, H. E. Stanley, and S. Havlin, “Local structure can identify and quantify influential global spreaders in large scale social networks,” *Proceedings of the National Academy of Sciences USA*, vol. 115, no. 29, pp. 7468–7472, 2018.

[17] L. Lu, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, “Vital nodes identification in complex networks,” *Physics Reports*, vol. 650, pp. 1 – 63, 2016.

[18] J.-X. Zhang, D.-B. Chen, Q. Dong, and Z.-D. Zhao, “Identifying a set of influential spreaders in complex networks,” *Scientific reports*, vol. 6, p. 27823, 2016.

[19] L. Lu, Y.-C. Zhang, C. H. Yeung, and T. Zhou, “Leaders in social networks, the delicious case,” *PloS one*, vol. 6, no. 6, p. e21202, 2011.

[20] E. Estrada and J. A. Rodriguez-Velazquez, “Subgraph centrality in complex networks,” *Physical Review E*, vol. 71, no. 5, p. 056103, 2005.

[21] D.-B. Chen, H. Gao, L. Lu, and T. Zhou, “Identifying influential nodes in large-scale directed networks: the role of clustering,” *PloS one*, vol. 8, no. 10, p. e74555, 2013.

[22] G. F. De Arruda, A. L. Barbieri, P. M. Rodriguez, F. A. Rodriguez, Y. Moreno, and L. da Fontoura Costa, “Role of centrality for the identification of influential spreaders in complex networks,” *Physical Review E*, vol. 90, no. 3, p. 032812, 2014.

[23] K. Klemm, M. Á. Serrano, V. M. Eguíluz, and M. San Miguel, “A measure of individual role in collective dynamics,” *Scientific reports*, vol. 2, p. 292, 2012.

[24] Supplementary file available here.

[25] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, “Epidemic processes in complex networks,” *Reviews of modern physics*, vol. 87, no. 3, p. 925, 2015.

[26] L. C. Freeman, “A set of measures of centrality based on betweenness,” *Sociometry*, pp. 35–41, 1977.

[27] G. Sabidussi, “The centrality index of a graph,” *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.

[28] P. Bonacich, “Factoring and weighting approaches to status scores and clique identification,” *Journal of mathematical sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[29] L. Katz, “A new status index derived from sociometric analysis,” *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[30] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[31] T. Martin, X. Zhang, and M. Newman, “Localization and centrality in networks,” *Physical review E*, vol. 90, no. 5, p. 052808, 2014.

[32] A. Braunstein, L. Dall’Asta, G. Semerjian, and L. Zdeborová, “Network dismantling,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 44, pp. 12368–12373, 2016.

[33] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, M. Muchnik, H. E. Stanley, and H. A. Makse, “Identification of influential spreaders in complex networks,” *Nature physics*, vol. 6, no. 11, p. 888, 2010.

[34] D. Chen, L. Lu, M.-S. Shang, Y.-C. Zhang, and T. Zhou, “Identifying influential nodes in complex networks,” *Physica a: Statistical mechanics and its applications*, vol. 391, no. 4, pp. 1777–1787, 2012.

[35] L. Lu, T. Zhou, Q.-M. Zhang, and H. E. Stanley, “The h-index of a network node and its relation to degree and coreness,” *Nature communications*, vol. 7, p. 10168, 2016.

[36] L. Zdeborová, P. Zhang, and H.-J. Zhou, “Fast and simple de-cycling and dismantling of networks,” *Scientific reports*, vol. 6, p. 37954, 2016.

[37] F. Moreone and H. A. Makse, “Influence maximization in complex networks through optimal percolation,” *Nature*, vol. 524, no. 7563, p. 65, 2015.

[38] P. Clausell, P. Grassberger, F. J. Pérez-Reche, and A. Politi, “Immunization and targeted destruction of networks using explosive percolation,” *Physical review letters*, vol. 117, no. 20, p. 208301, 2016.

[39] F. Radicchi and C. Castellano, “Leveraging percolation theory to single out influential spreaders in networks,” *Physical Review E*, vol. 93, no. 6, p. 062314, 2016.

[40] F. Radicchi and C. Castellano, “Fundamental difference between superblockers and superspreaders in networks,” *Physical Review E*, vol. 95, no. 1, p. 012318, 2017.

[41] Supplementary file available here.

[42] R. Guimerà, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, “Self-similar community structure in a network of human interactions,” *Physical review E*, vol. 68, no. 6, p. 065103, 2003.

[43] M. Newman and R. Ziff, “Efficient monte carlo algorithm and high-precision results for percolation,” *Physical Review Letters*, vol. 85, no. 19, p. 4104, 2000.
[44] F. Radicchi, “Predicting percolation thresholds in networks,” *Physical Review E*, vol. 91, no. 1, p. 010801, 2015.
[45] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *science*, vol. 286, no. 5439, pp. 509–512, 1999.
[46] V. Gómez, A. Kaltenbrunner, and V. López, “Statistical analysis of the social network and discussion threads in slashdot,” in *Proceedings of the 17th international conference on World Wide Web*, pp. 645–654, ACM, 2008.
[47] J. Kunegis, “KONECT – The Koblenz Network Collection,” in *Proc. Int. Conf. on World Wide Web Companion*, pp. 1343–1350, 2013.
[48] M. Ripeanu, I. Foster, and A. Iamnitchi, “Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design,” *arXiv preprint arXiv:0209028*, 2002.
[49] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graph evolution: Densification and shrinking diameters,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 2, 2007.
[50] M. Richardson, R. Agrawal, and P. Domingos, “Trust management for the semantic web,” in *The Semantic Web-ISWC 2003*, pp. 351–368, Springer, 2003.
[51] J. McAuley and J. Leskovec, “Learning to discover social circles in ego networks,” in *Advances in Neural Information Processing Systems*, pp. 548–556, 2012.
[52] E. Cho, S. A. Myers, and J. Leskovec, “Friendship and mobility: user movement in location-based social networks,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1082–1090, ACM, 2011.
[53] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, “Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters,” *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.
[54] J. Leskovec, L. A. Adamic, and B. A. Huberman, “The dynamics of viral marketing,” *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, p. 5, 2007.
[55] J. Yang and J. Leskovec, “Defining and Evaluating Network Communities based on Ground-truth,” in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, p. 3, ACM, 2012.