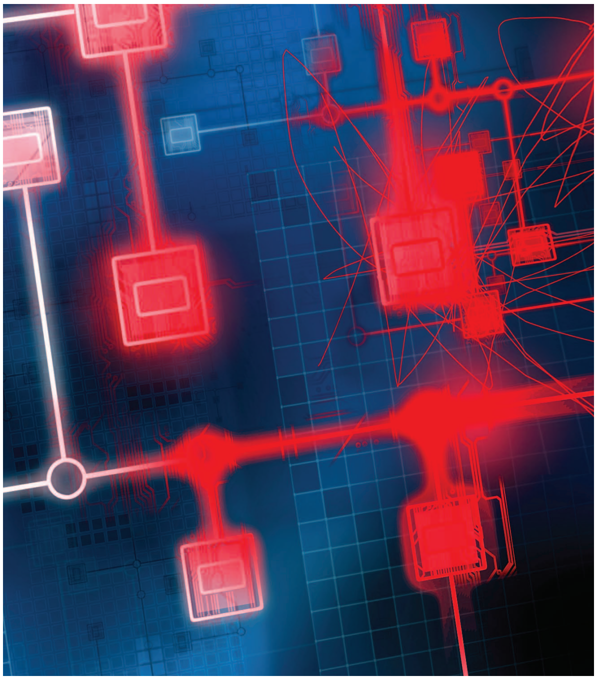

Detecting Influential Spreaders in Complex, Dynamic Networks

Pavlos Basaras, Dimitrios Katsaros, and Leandros Tassiulas, *University of Thessaly, Greece*

A hybrid of node degree and *k*-shell index is more effective at identifying influential spreaders and has less computational overhead than either of these traditional measures.

With the unprecedented growth during the past decade of different types of social and enterprise networks, alongside naturally occurring networks in human communities, society is on the verge of becoming "fully networked."

Recent advances in information and communications technologies, coupled with the ability to create and store a vast amount of data on various aspects of human behavior, have made it possible to analyze complex networks. Studies range from purely graph-theoretic aspects (size and strength of communities, robustness to attacks, growth models, node connectivity, and so on), to more social-theoretic aspects (for example, homophily and rumor spreading). This research has given rise to computational social science,[1] a new field that leverages the ability to collect and analyze data to reveal hidden patterns in individual and group activities.

Insights into complex networks' structural and topological properties have informed work in numerous areas including search engine technology,[2] the development of ad hoc network protocols,[3] and detecting and containing disease outbreaks.[4] Security researchers have likewise used complex network analysis to study terrorist networks,[5] virus propagation over computer networks, and resistance to cyberattacks. Such analyses typically apply graph theory and involve centrality measures, shortest-path algorithms, degree distributions, and so on.

Here, we focus on the problem of *influential spreaders*—nodes in complex networks that can spread a message rapidly among other nodes. Early detection of such entities can help security technologists prevent extended damage to networks against malware or, in the case of terrorist networks, identify the most important malefactors.

To identify influential spreaders, researchers traditionally have relied on the k-*shell index*,[6] a degree-based measure of a node's "coreness." However, the significant computational overhead of this index makes it inappropriate for analyzing dynamic networks.

We propose an alternative measure, the *μ-power community index*, that is an amalgam of coreness and betweenness centrality; μ-PCI is calculated in a completely localized manner and thus suitable for any kind of network irrespective of its size or dynamicity.[3] An experimental evaluation of the two values, along with a baseline measure based solely on node degree, demonstrates μ-PCI's superiority in detecting influential spreaders.

MOTIVATION

Consider an example in which an attacker installs a virus on a host mobile device with the intention of exploiting the host's connections to spread the malware

Published by the IEEE Computer Society

and ultimately infect as many other devices as possible. Assume that all devices comprise a single network with common administration. Upon detecting the malware, the administrator immediately takes action to limit its propagation. Possible measures include installing more effective antivirus software to selected devices, shutting these devices down, or disconnecting them from the rest of the network.

Two well-known cases of malware that exploit mobile devices' network connections are the Cabir and Commwarrior-A worms. The former spreads through Bluetooth connections to other Bluetooth-enabled devices that it can find. The latter was the first worm to propagate via the Multimedia Messaging Service; it searches through a user's local address book for phone numbers and sends MMS messages containing infected files to other users.

Obviously, if the infected devices in our scenario are influential spreaders, they will impact a large part of the network. This leads to several questions: How fast will the virus spread? Is the infection rate different in different network topologies? Does the percentage of infected nodes in the network depend on the node(s) where the infection originated? Do multiple infection starting points produce a substantially broader infection area? If so, what does this depend on? Which nodes should the administrator disconnect to stop the propagation?

Researchers who have investigated such questions found that not all nodes in a complex network have the same potential to propagate a message efficiently.[6,7] Explanations for this behavior range from a network's topological characteristics at global scale—for example, power-law degree connectivity—to individual nodes' connectivity patterns.

IDENTIFYING INFLUENTIAL SPREADERS

Most studies of influential spreaders have focused on their linkage with other nodes. The problem has not been described formally but is similar to two others: detecting a network's central nodes and selecting the set of nodes that maximize the spread of infection.

Identifying the central nodes in a complex network usually relies on graph-theoretic concepts of *betweenness centrality*. Such measures are generally based on a node's degree or on its geodesic distance to other nodes.[8] The former category includes degree centrality, spectral centrality, and coreness, whereas the latter includes closeness, shortest-path, and bridging centrality. Degree-based centrality measures consider a node prominent if its connections make it visible to the network's other nodes. Intuitively, a node is prominent if it is adjacent to many other prominent nodes. The latter family of centrality measures exploits the shortest path between nodes.

The spread maximization problem has been proved to be NP-hard in threshold networks,[9] and researchers have proposed several greedy algorithms to solve it—for example, there are simple and efficient algorithms that adopt the voter model.

Recent studies of social networks have considered other node features besides connectivity such as age, gender, and marital status.[10] Another feature is trustworthiness, which can affect a decision to follow a link to malware. Examples of malware that exploited trust to spread across a social network include the Skype and Koobface worms.

BALANCING BETWEENNESS AND CORENESS

Maksim Kitsak and his colleagues found that the degree of a node is not a good indicator of its ability to spread a message to a sufficiently large part of the network.[6] They reported that measures based on betweenness centrality are distorted by the degree-1 node, which increases the centrality index of the sole node connected to them.

Our own research found that exploiting betweenness centrality has several disadvantages for disseminating messages in wireless ad hoc networks.[3] Relying on a degree-1 node results in overestimating the spreading capabilities of a node connected to it. Moreover, based on a detailed investigation of the spreading capabilities of high-degree nodes in various complex networks, we found that high-degree nodes are indeed often good spreaders.

Exploiting betweenness centrality has several disadvantages for disseminating messages in wireless ad hoc networks.

Kitsak's team argued that the node's position in a *k*-shell decomposition of the network's graph is a better way of quantifying influential spreaders, and went on to verify this hypothesis in the context of disease propagation.[6] However, subsequent research proved that a node's spreading capabilities in the context of rumor spreading do not depend on its *k*-core index.[11]

As the "*K*-Shell Decomposition" sidebar explains, this approach has two other major shortcomings. First, it has significant computational overhead, rendering it unsuitable for dynamic networks. Second, it is impossible to guarantee a monotonic relationship between the *k*-shell index and a node's spreading capability, which causes major problems when there are not enough resources to expend on node vaccination.

We have developed a method that quantifies spreading capabilities in a completely localized manner, making it suitable for any kind of network, irrespective of size or dynamicity.[3] This metric, μ-PCI, balances the principles of betweenness centrality—it considers nodes that lie on many communicating paths between pairs of nodes—and the transitive network density implied by the coreness measure. The metric is computed as follows: the μ-PCI of a node v is equal to k, such that there are up to $\mu \times k$ nodes in the μ-hop neighborhood of v with degree greater than or equal to k, and the rest of the nodes in that neighborhood have a degree less than or equal to k. The goal is to detect nodes located in dense areas of the network and thus likely influential spreaders.

K-Shell Decomposition

K-shell decomposition of a network graph is performed iteratively. The first step involves removing all degree-1 nodes, along with their link, and indexing these as $k = 1$. In the resulting graph, all nodes of degree 1 are also considered to have $k = 1$ and are again pruned. The process is repeated until there are no nodes of degree 1. Similarly, all nodes with i or fewer connections are iteratively removed; these nodes are indexed as $k = i$.

The *k*-shell method has two significant disadvantages for defining influential spreaders.

First, although *k*-shell decomposition is an easy task from an algorithmic perspective, the measure itself is not localized; hence, determining the *k*-shell index requires both global knowledge of the network topology and multiple iterations. Although a recent attempt to implement *k*-shell decomposition in a distributed manner achieved an 80 percent reduction in execution time, the researchers offered no alternative to the algorithm's iterative nature.[1] Thus, this solution cannot be applied in contexts with real-time requirements, such as security applications.

Second, *k*-shell decomposition frequently fails to establish a monotonic relation between k and the total infection area, which could have severe implications. An administrator who has limited time (m actions) or resources to shut down some uninfected machines would prefer to select those with the maximal spreading capabilities. If there were a strictly monotonic relation between k and the total infection area, the administrator would choose m nodes among those with the maximal *k*-shell indices. Unfortunately, in many cases *k*-shell decomposition provides no such guarantees; quite often nodes with maximum k do not offer high enough spreading capabilities. Hence, a desirable property is for a measure to violate monotonicity as rarely as possible.

Reference

1. A. Montresor, F. De Pellegrini, and D. Miorandi, "Distributed K-Core Decomposition," *IEEE Trans. Parallel and Distributed Systems*, vol. 24, no. 2, 2013, pp. 288-300.

PERFORMANCE EVALUATION

To evaluate our technique's accuracy, we compared it to k-shell decomposition and a baseline measure based solely on the node degree on a large number of complex networks. Here, we present the most significant findings from ca-CondMat and ca-AstroPh—two well-known collaboration networks from the e-print arXiv covering condensed matter physics and astrophysics, respectively—from the Stanford Network Analysis Platform (http://snap.stanford.edu/data). Table 1 summarizes the networks' main characteristics.

Table 1. Complex network attributes.

Network	Type	No. of nodes	No. of links	Infection probability (%)
ca-CondMat	Sparse	23,133	186,936	8
ca-AstroPh	Dense	18,772	396,160	4

We used the susceptible-infected-recovered model for an infection originating from both a single spreader and multiple spreaders to investigate the spreading process, as detailed in the "Infection Origins" sidebar. SIR models three possible states:

- the susceptible state S, in which the S nodes are vulnerable to infection;
- the infected state I, in which the I nodes try to infect their susceptible neighbors and succeed with probability λ; and
- the recovered state R, in which nodes have recovered from infection and cannot be reinfected.

We used relatively small values of λ to highlight the role of influential spreaders.

We compared μ-PCI, *k*-shell decomposition, and the node degree method. For μ-PCI, we present only results for $\mu = 1$. We obtained analogous results for $\mu = 2$, but the method's performance deteriorates substantially for $\mu > 2$. We use *km*, *ks*, and *k* to represent the 1-PCI, *k*-shell index, and node degree values, respectively.

Similar to Kitsak and his colleagues,[6] we used the average size of the network's infected area as a performance measure. To quantify $inf(s)$, the influence of a single spreader s, we computed the average size of the network infected with the (km, k) pair values. We averaged the extent of the infected network over all spreaders as follows:

$$INF_{km,k} = \sum_{s \in P_{km,k}} \frac{inf(s)}{N_{km,k}},$$

where $P_{km,k}$ is the set of all $N_{km,k}$ spreaders with the same (km, k). We repeated the same process for *k*-shell decomposition.

To obtain statistically unbiased results, we repeated the computation 1,000 times for each vertex of a graph for the single- and multiple-origin scenarios.

We found that 1-PCI exhibits steady and reliable behavior, overcoming the disadvantages of high-degree spreaders and of *k*-shell decomposition. Choosing high 1-PCI nodes maximizes spreading influence, whereas selecting the high-degree nodes or a random node from the core shell either results in poor spreading or does not maximize influence.

Single original spreader

Our first experiment examined the three methods' ability to select the most influential spreaders for a single-origin process.

Figure 1 shows all nodes' spreading capability in the ca-CondMat network according to their 1-PCIs and *k*-shell indices. The 1-PCI method results in a more monotonic distribution than *k*-shell decomposition, providing a clearer ranking of spreading capabilities. It converges to an approximately straight line, where maximum influence lies, more steeply than the *k*-shell method in all the studied cases. Choosing a spreader with, say, 1-PCI > 23, will yield the maximum influence, whereas choosing one from the core or from the high shells might not be optimal because in some cases nodes within the same shell have different spreading capabilities.

Figure 2 shows all nodes' spreading capability in the ca-AstroPh network according to their 1-PCIs and *k*-shell indices versus the respective node's degree. In particular, the plots depict the average size of the infected population $INF_{km,k}$ for all spreaders with (1-PCI, degree) pair values. The *k*-shell index clearly fails to fulfill monotonicity in many cases. Also, 1-PCI has a better correlation with node degree.

This experiment confirmed the conclusion of Kitsak's team that measures such as node degree cannot accurately predict a network's most influential spreaders.[6] For a fixed degree equal to k, there is a wide spectrum of $INF_{km,k}$ values, making the degree measure an ineffective solution,

Infection Origins

Our performance evaluation considered infections originating with both a single spreader and multiple spreaders.

SINGLE ORIGINAL SPREADER

All nodes are initially at the susceptible (*S*) state, except for one node, which is in the infected (*I*) state. The infected node tries to infect its susceptible neighbors with probability of success λ, and then changes to the recovered (*R*) state. All nodes in state *I* try to infect their susceptible neighbors, and the process repeats until there is no node in the *I* state.

MULTIPLE ORIGINAL SPREADERS

The number of initially infected nodes ranges from 0.5 to 4 percent of the network's total size.

μ-PCI and node degree methods

The malicious set of spreaders is empty in the first phase. We introduce the spreader with the highest value of each method to its respective set; we then select the spreader with the next highest value, which is not connected to the previous set. The process repeats until the initial infection percentage of the network is satisfied.

K-shell decomposition

Because all spreaders in each shell are treated evenly, we start by introducing a randomly selected node to the set. We randomly select the next spreader from the remaining nodes of the core shell that are not directly connected to the previous set, and continue this process iteratively. If the initial infection percentage cannot be met from the core shell, we repeat the process on the shell immediately below it, and so on.

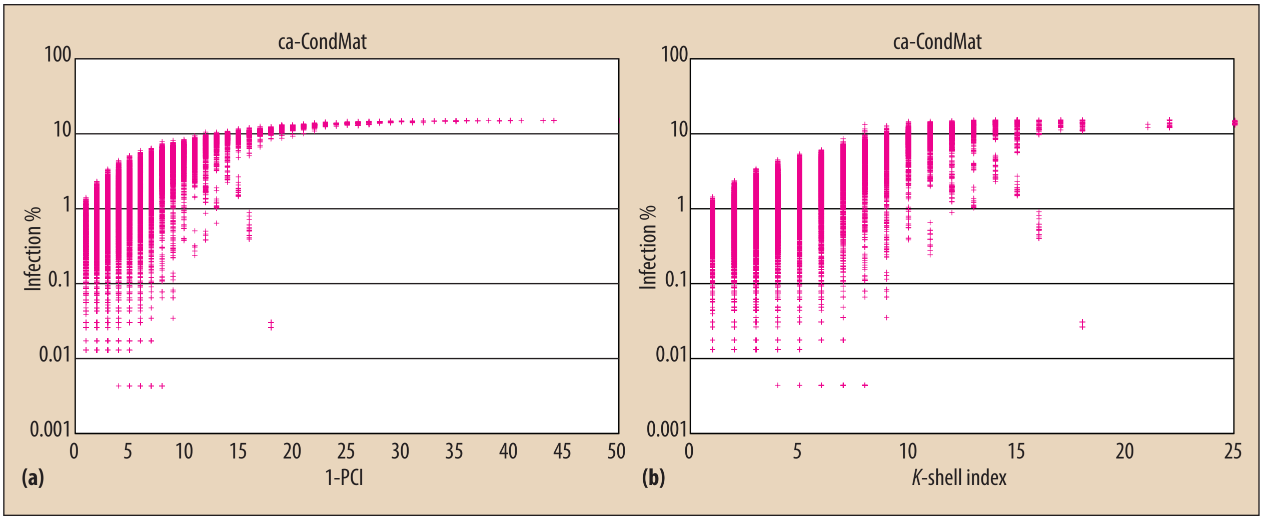

Figure 1. Spreading capability of nodes in the ca-CondMat network with a single original spreader according to (a) 1-PCI and (b) *k*-shell index. There are nodes with high *k*-shell indices, some of which infect a large portion of the network, as well as nodes with the same *k*-shell index (16) that infect a significantly smaller part of the network. On the other hand, only nodes with very small 1-PCI exhibit such behavior.

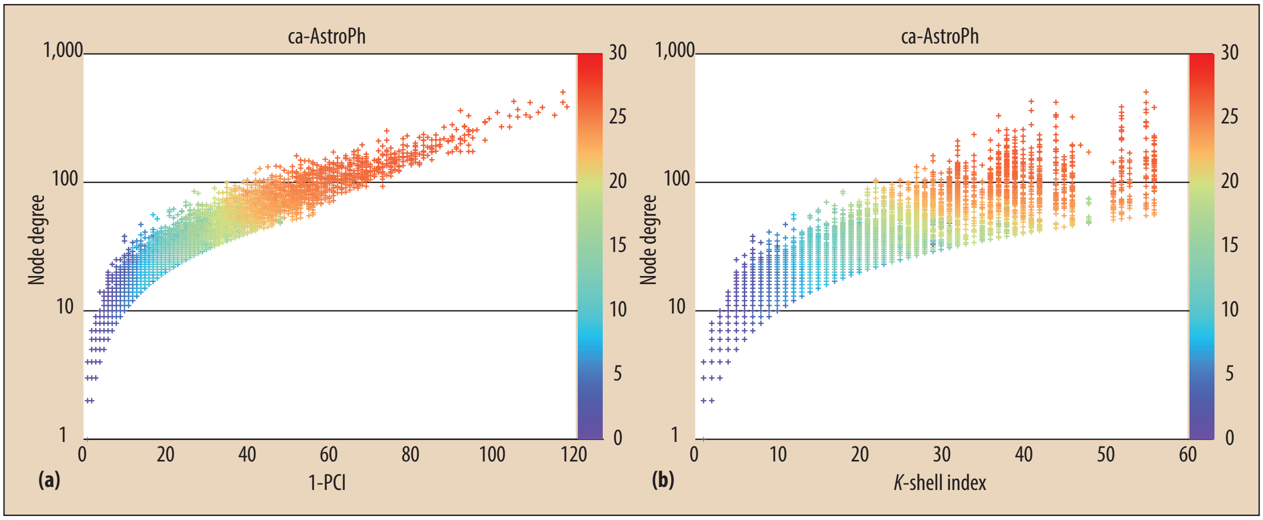

Figure 2. Spreading capability of nodes in the ca-AstroPh network with a single original spreader according to (a) 1-PCI and (b) *k*-shell index versus node degree. The *k*-shell index fails to fulfill monotonicity in many cases, and 1-PCI has a better correlation with node degree.

Table 2. Number of influential spreaders that can maximize infection in three networks.

Network	Size (nodes)	Density (edges)	Infected area (%)	No. of influential spreaders
soc-Slashdoc0811	77,360	905,468	16.5	1,788
ca-CondMat	23,133	186,936	1.9	127
ca-AstroPh	18,772	396,160	26.5	477

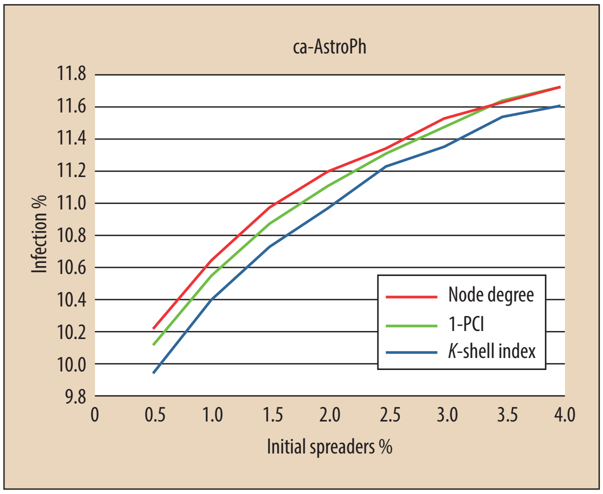

Figure 3. Spreading capability of nodes in the ca-AstroPh network with multiple original spreaders according to node degree, 1-PCI, and *k*-shell index. The *k*-shell index is the least effective measure. Node degree is the most effective measure, closely followed by 1-PCI, but the discrepancy between these values quickly diminishes as the number of multiple original spreaders grows.

especially in cases where the objective is to select a very small number of spreaders. This occurs because a high-degree node might be located in a sparse neighborhood.

The *k*-shell index depends less on node degree when moving to higher shells, but the best spreaders are often scattered across numerous shells, thus violating monotonicity. Most nodes have a *k*-shell index equal to 48, which is particularly high; their spreading capability is similar to that of nodes with a *k*-shell value less than 30.

For a fixed 1-PCI, the infection percentage is approximately the same and independent of node degree, making high 1-PCI nodes the best choice in single-origin spreading processes. The 1-PCI measure groups spreaders according to their spreading capabilities: lower 1-PCI values correspond to poor spreaders, whereas high values indicate the most influential ones.

As a node's 1-PCI increases, its spreading influence also appears to increase. Consider, for example, the results obtained from the ca-AstroPh network shown in Figure 2b. Moving to higher shells—starting at, say, $ks > 34$—spreading influence seems to constantly increase. However, this increase stops at $ks = 48$, where the infection decreases drastically. The 1-PCI analysis does not elicit such behavior, especially when close to maximum influence. As 1-PCI values increase, influence also continuously increases until maximum infection is reached.

We computed the number of influential spreaders that can achieve the maximum infection (with 1 percent deviation) for the two networks described here along with the soc-Slashdoc0811 network. As Table 2 shows, network size and topology impact the number of influential spreaders. We observed no increasing or decreasing relation between the number of influential spreaders and network size—the key factor is the pattern of node connections.

Multiple original spreaders

Our second experiment examined the three methods' ability to select the most influential spreaders for a multiple-origin process. To maximize the infected area, the original spreaders were not linked. If selected spreaders were connected, the infected region would be smaller due to the overlap of neighboring spreaders' "influence regions."[6]

Figure 3 shows all nodes' spreading capability in the ca-AstroPh network according to their degree, 1-PCI, and *k*-shell index. The *x*-axis indicates the percentage of initially infected nodes, with λ at 2 percent. The results were similar for other networks.

Although high 1-PCI nodes are the most influential spreaders in a single-origin process, all three measures are comparable in this case. The *k*-shell index is the least effective measure. Node degree is the most effective, closely followed by 1-PCI, but the discrepancy between these values quickly diminishes as the number of multiple original spreaders grows.

Discovering the most influential spreaders is the key to immunizing complex, dynamic networks against cyberattacks and thereby limiting infection. Overall, μ-PCI, which can be considered a hybrid of node degree and *k*-shell index, is more effective at identifying influential spreaders and has less computational overhead than either of these traditional measures. Further work could include the use of control-theoretic techniques to improve results. C

Acknowledgments

We thank the anonymous reviewers and Christos Kapoutsis for their valuable comments and suggestions. This work has been supported by the EU project "REDUCTION" (no. 288254), and by the EU Network of Excellence "Internet Science" (no. 288021).

References

1. D. Lazer et al., "Computational Social Science," *Science*, 6 Feb. 2009, pp. 721-723.
2. A.N. Langville and C.D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton Univ. Press, 2006.
3. D. Katsaros, N. Dimokas, and L. Tassiulas, "Social Network Analysis Concepts in the Design of Wireless Ad Hoc Network Protocols," *IEEE Network*, vol. 24, no. 6, 2010, pp. 23-29.
4. N.A. Christakis and J.H. Fowler, "Social Network Sensors for Early Detection of Contagious Outbreaks," *PLOS ONE*, vol. 5, no. 9, 2010; doi:10.1371/journal.pone.0012948.
5. V.E. Krebs, "Uncloaking Terrorist Networks," *First Monday*, vol. 7, no. 4, 2002; http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/941/863.
6. M. Kitsak et al., "Identification of Influential Spreaders in Complex Networks," *Nature Physics*, vol. 6, 2010, pp. 888-893.
7. B. Doerr, M. Fouz, and T. Friedrich, "Why Rumors Spread So Quickly in Social Networks," *Comm. ACM*, vol. 55, no. 6, 2012, pp. 70-75.
8. L.C. Freeman, "Centrality in Social Networks: Conceptual Clarification," *Social Networks*, vol. 1, no. 3, 1979, pp. 215-239.
9. D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the Spread of Influence through a Social Network," *Proc. 9th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (KDD 03), ACM, 2003, pp. 137-146.
10. S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," *Science*, 21 June 2012, pp. 337-341.
11. J. Borge-Holthoefer and Y. Moreno, "Absence of Influential Spreaders in Rumor Dynamics," *Physical Rev. E*, vol. 85, no. 2, 2012; doi:10.1103/PhysRevE.85.026116.

Pavlos Basaras *is a PhD student in the Department of Computer & Communications Engineering at the University of Thessaly, Greece. His research focuses on complex network analysis. Basaras received an MSc in computer engineering from the University of Thessaly. Contact him at pbasaras@gmail.com.*

Dimitrios Katsaros *is a lecturer in the Department of Computer & Communications Engineering at the University of Thessaly, Greece. His research interests include distributed systems and complex networks. Katsaros received a PhD in informatics from the Aristotle University of Thessaloniki, Greece. Contact him at dkatsar@inf.uth.gr.*

Leandros Tassiulas *is a professor in the Department of Computer & Communications Engineering at the University of Thessaly, Greece. His research interests include computer and communication networks. Tassiulas received a PhD in electrical engineering from the University of Maryland, College Park. He is an IEEE Fellow. Contact him at leandros@uth.gr.*