# Towards Generalized Centrality Measures with Applications to Information Networks

Evimaria Terzi & Azer Bestavros & Dora Erdos & Vatche Ishakian
Computer Science Department, Boston University, Massachusetts

The problem of identifying "valuable" or "central" nodes in a given network has long been recognized as important by researchers and practitioners alike. There exists an abundance of measures that associate each node with an individual centrality score; the higher the score of a node the more central its position in the network. Link-analysis algorithms and node-centrality measures try to capture this intuition [1, 5, 6, 8, 10]. While existing, commonly used centrality measures give intuition about the relative value of individual nodes, they are not useful in assessing the *collective value of groups of nodes*, which is not necessarily nor typically reflected by the sum of the values of the nodes in the group. For example, assume that the centrality of a group of nodes is defined as the total number of shortest paths passing through at least one node in the group. In this case, a set of nodes with high group centrality may not necessarily include nodes with high individual centrality scores. A relatively small number of recent studies considered such combinatorial notions of nodes' importance in conjunction with specific problems, including advertisement strategy design [4], virus containment [7], and shortest-path distance approximation [9].

The goal of this paper, is to present the blueprints of a research agenda, which explores expressive notions of *group centrality* and develops algorithmic techniques, which enable the implementation and evaluation at scale of an arsenal of tools for use by researchers and practitioners. Next, we present some indicative applications, which guide and will also benefit from the development of this agenda.

**Content de-duplication in information-flow networks:** Consider flow networks where data items propagate to the network nodes. Examples of such data items include updates in social networks, news flowing through interconnected RSS feeds and blogs, measurements in sensor networks, route updates in ad-hoc networks. Oftentimes, such propagation lacks coordination: nodes relay information they receive to neighbors, independent of whether or not these neighbors received the same information from other sources. This uncoordinated data dissemination may result in significant, yet unnecessary, communication and processing overheads, ultimately reducing the utility of information networks. To alleviate the negative impacts of this *information multiplicity* phenomenon, we propose that a subset of nodes, that we call *filters*, carry out additional information de-duplication functionality. The strategic placement of filters will determine the extent of information multiplicity and ultimately the level of user satisfaction. In this context, the central nodes correspond to the selected filters. Observe that the placement of the filters does not affect the information that nodes receive; it only reduces the multiplicity of received copies. Our preliminary work [2] indicates that this problem is NP-hard, however efficient approximation algorithms exist. Knowledge of the network structure as well as network-sampling methods can further benefit the performance of these algorithms so that they also handle large datasets.

**Information gathering in information-flow networks:** Consider the problem of identifying the minimum set of (or best fixed number of) nodes to use for capturing all of (or most of) the information propagating through a flow network from a set of sources to a set of destinations. Since information does not propagate in such networks through shortest (or even single) paths, this problem reduces to the identification of the set of nodes that collectively lie on all (or most) paths in the network. Although existing applications dictate such centrality definitions, the computational complexity of the task of finding the set of such nodes is extremely high – after all, there are expo-

nentially many paths! In our recent work [3], we studied how the computational complexity of this problem is affected by the structure of the underlying flow network graph (e.g., tree, acyclic graph etc). For many of these cases were polynomial (approximation) algorithms exist it is interesting to explore the type and the power of accurate sampling techniques.

**Effective advertising strategies in navigational networks:** The analysis of navigational patterns – governed by an underlying access network – is instrumental for identifying the set of nodes on which to place an advertisement (ad) for maximal exposure. Studies have shown that the number of times a person is exposed to an ad in a short period of time correlates with response probability; this number is known as the *effective frequency* (`http://en.wikipedia.org/wiki/Effective_frequency`). In this context, an interesting problem is the following: assuming an effective frequency of $\ell$, what set of $k$ nodes in an access network should be selected for ad placement? Observe that the effective-frequency parameter requires that the ad messages be placed so that users encounter them in (almost) consecutive pages in a given browsing session. In fact, we can extend such centrality definition even further: instead of requiring ad messages to be placed on a set of neighboring nodes, we can impose the requirement that they are placed on strongly connected subgraphs of the underlying access network. Both of these formulations (as well as others that we cannot fully present due to space limitations) give rise to new algorithmic challenges, and – perhaps more importantly from a broader impact perspective – to new types of advertisement strategies.

All the examples presented above, give rise to new combinatorial notions of centrality. The ability to solve such problems is useful both for researchers and practitioners. We believe that the development of centrality-as-a-service is an interesting and important direction for this line of research. Taking bibliography data as an example, one can develop a tool that finds central authors or central papers within a particular scientific domain.

## References


[1] S. Brin, R. Motwani, L. Page, and T. Winograd. What can you do with a web in your pocket? *IEEE Data Eng. Bull.*, 21(2):37–47, 1998.

[2] D. Erdos, V. Ishakian, A. Lapets, E. Terzi, and A. Bestavros. The filter-placement problem and its application to minimizing information multiplicity. In *International Conference on Very Large Databases (VLDB)*, 2012.

[3] V. Ishakian, D. Erdos, E. Terzi, and A. Bestavros. A framework for the evaluation and management of network centrality. In *Siam Data Mining Conference (SDM)*, 2012.

[4] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[5] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA*, pages 668–677, 1998.

[6] R. Lempel and S. Moran. Salsa: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, 2001.

[7] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. M. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007.

[8] M. E. J. Newman. The mathematics of networks. *The New Palgrave Encyclopedia of Economics*, 2008.

[9] M. Potamias, F. Bonchi, C. Castillo, and A. Gionis. Fast shortest path distance estimation in large networks. In *CIKM*, pages 867–876, 2009.

[10] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.