Workshop on Incomplete Network Data Held at Sandia National Labs – Livermore
By Sucheta Soundarajan (Syracuse University) and Jeremy D. Wendt (Sandia National Laboratories)

While network analysis is applied in a broad variety of scientific fields (including physics, computer science, biology, and the social sciences), how networks are constructed and the resulting bias and incompleteness have drawn more limited attention. For example, in biology, gene networks are typically developed via experiment -- many actual interactions are likely yet to be discovered. In addition to this incompleteness, the data-collection processes can introduce significant bias into the observed network datasets [1][2]. For instance, if you observe part of the World Wide Web network through a classic random walk, then high degree nodes are more likely to be found than if you had selected nodes at random [3]. Unfortunately, such incomplete and biasing data collection methods must be often used.

At the recent Workshop on Incomplete Network Data (WIND), held at Sandia National Laboratories[1] in Livermore, California, researchers from academia, industry, and national labs gathered to discuss perspectives on dealing with incomplete network data. WIND was organized by Tina Eliassi-Rad (Northeastern University), James Ferry (Metron, Inc.), Ali Pinar (Sandia), and C. Seshadhri (University of California, Santa Cruz). The complete schedule is available on-line (http://eliassi.org/WIND16.html). Thus, references to WIND talks refer only to the speaker's name and affiliation.

A host of areas with biased graph samples were discussed at WIND. Dennis Feehan (University of California, Berkeley) and Forrest Crawford (Yale University) both discussed a particularly interesting problem from the social sciences -- determining how to accurately estimate the size of hidden or rare groups in massive populations by querying survey respondents. Bradley Huffaker (Center for Applied Internet Data Analysis, University of California, San Diego) presented problems related to obtaining an accurate map of the Internet, and Jaiwei Han (University of Illinois at Urbana-Champaign) presented techniques for supplementing explicit graphs using unstructured text mining.

Three main categories of approaches emerged. The first approach was estimating properties or characteristics about the global network given only a partial observation of that network. For example, given only partial access to the full network data, can one estimate the number of triangles (i.e., "A" knows "B" knows "C" knows "A") in the full network (e.g., Tammy Kolda - Sandia). The second approach is performing data collection in such a way as to reduce bias or increase the quality of the information obtained. For instance, how can one sample a node

---

[1] Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

from the graph uniformly at random from the graph, where access to data is through a random walk-like crawl (e.g., Ravi Kumar - Google)? The third approach is identifying algorithm degradation resulting from noise or incomplete data and designing algorithms to be more robust. For example, local spectral methods provide results akin to full-graph spectral methods, but without being affected by problems in distant parts of the graph (e.g., Michael Mahoney – University of California, Berkeley; David Gleich – Purdue University).

These categories are complementary, and a real-world application of network analysis on incomplete data would ideally incorporate all three techniques. The third category (understanding the robustness of existing algorithms against noise or incompleteness) showed itself to be an important first step in any such unified approach.

At WIND, David Kempe (University of Southern California) discussed the problem of algorithm robustness. He pointed out that in real-world network data, "Noise is the norm, not the exception," and that understanding the effects of noise on algorithmic tasks is critical. To illustrate this point, he considered the problem of influence maximization: In a network setting, if we assume that an individual's beliefs can affect their neighbors' beliefs, which nodes' beliefs should we influence to have the greatest effect on the beliefs of the population as a whole? Kempe argued that the influence probabilities (i.e., the probability that node "A" will influence the belief of node "B") can have a large effect on which nodes are selected; but any estimates of these probabilities are likely to be inaccurate!

Kempe also considered the effect of noise or incomplete data on community detection (the problem of clustering the nodes of a network into cohesive groups). He argued that the output of community detection methods can also be significantly affected by noise or missing edges in the network. For example, missing edges might lead an algorithm to identify two communities, while if those edges had been present, it would have found only one community. Kempe argued that community detection on incomplete network datasets may be appropriate for suggesting hypotheses, which are then verified by other means, but not for drawing conclusions.

Along similar lines, Anil Vullikanti (Virginia Tech) considered how noise can affect the core decomposition of a graph. A core of a graph is, in essence, a 'dense' or 'central' part of the graph and, among other applications, can be used to measure the importance or centrality of nodes in the network. Through experimental results, Vullikanti demonstrated that k-cores are unstable when the network is perturbed in degree-biased ways (that is, the probability of a perturbation affecting a node depends on the number of connections that the node has). This is a critical problem because one of the most common ways of obtaining network data (crawl via breadth-first search) leads to just this kind of degree-biased sampling.

Other research presented during the workshop suggested techniques to overcome missing data or noise, including strategies for counteracting bias or generating more accurate network samples. The consensus among the WIND attendees was that incomplete data presents a daunting challenge to performing accurate network analysis. Several attendees presented early solutions that show great promise. However, several critical questions remain: How do we measure or estimate the noise, bias, or incompleteness of network datasets? What tests could we run to thoroughly test the effects of these data errors on later analyses?

1. J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *12th International Conference on Knowledge Discovery and Data Mining (KDD)*, Philadelphia, PA, August 20–23 2006.

2. A. S. Maiya and T. Y. Berger-Wolf, "Sampling community structure," in *19th International World Wide Web Conference*, Rayleigh, NC, April 26–30 2010.

3. M. Kurant, A. Markopoulou, and P. Thiran, "On the bias of BFS," in *22nd International Teletraffic Congress*, 2010, pp. 1–8.