Network Analytics in the age of Big Data

Nataša Przulj* and Noël Malod-Dognin

Department of Computer Science, University College London, UK

*To whom correspondence should be addressed; natasa@cs.ucl.ac.uk

We live in a complex world of inter-connected entities. In all areas of human endeavor, from biology to medicine, economics and climate science, we are flooded with large-scale datasets. They describe complex, real-world systems from different and complementary viewpoints, with entities being modeled as nodes and their connections as edges, comprising large networks. This is a new and rich source of domain-specific information, but that information is currently largely hidden from us within the complex wiring patterns. Deciphering them is a foremost scientific problem, because from theoretical computer science we know that computational analyses of large networks are often intractable, meaning that many of the questions we ask about our world we cannot answer exactly, even if we had all compute power and all the time of the universe [1]. Hence, our only hope is to answer them approximately (also called *heuristically*) and prove how far the approximate answer is from the exact, unknown one, in the worst case. On page XXX of this issue, Benson *et al.* [2] make an important step in that direction by providing a scalable heuristic framework for grouping entities based on their wiring patterns and utilizing the discovered patterns for revealing the higher-order organizational principles of several real-world networked systems.

To mine the wiring patterns of networked data and uncover the functional organization, it is not enough to consider only simple descriptors, such as the number of interactions that each entity (node) has with other entities (called *node degree*), because two networks can be identical in such simple descriptors, but have a very different connectivity structure (Fig. 1). Instead, Benson *et al.* [2] use higher-order descriptors based on small sub-networks obtained on a subset of nodes in the data that contain all interactions that appear in the data, called *graphlets* (e.g., a triangle) [3]. They identify network regions rich in instances of a particular graphlet type, with few of the instances of the particular graphlet crossing the boundaries of the regions. If the graphlet type is specified in advance, the method can uncover the nodes interconnected by it, which enabled Benson *et al.* to find 20 neurons that control a particular type of movement in the nematode worm neuronal network. In this way, the method unifies the local wiring patterning with higher order structural modularity imposed by it, uncovering higher order functional regions in networked data.

The importance of this result is in that it can be applied to a broad range of networked data whose understanding is fundamental to answering foremost questions humanity is facing today, from climate change and impacts of genetically modified organisms to the environment [4], to food security, human migrations, economic and societal crises [3,5], understanding diseases, aging and personalizing medical treatments [6-13]. For example, the cell is a complex system of interacting molecules, in which *genes* are transcribed into *RNAs* and translated into *proteins*, which bind together to change their 3-dimensional structure and do particular cellular functions. We capture various molecular interactions by different high-throughput biotechnologies and we model them with different types of networks. Individual analyses of molecular networks have revealed that molecules involved in similar functions tend to group together in a network and are similarly wired [14], which is improving our understanding of gene functions [6], molecular organization of the cell [7] and is impacting therapeutics [8-12].

However, each network type provides limited information about the phenomenon at study. For example, a disease is rarely the consequence of a single mutated gene, or of a single broken molecular interaction, but of multiple perturbations of complex interactions within and across cells. *Network Medicine* couples network analytics with data integration to mine the wealth of complementary data and reveal common molecular mechanisms between seemingly unrelated diseases [8-11]. In contrast, patients with seemingly the same disease may have very different molecular mechanisms of disease and reactions to treatment (e.g., cancer heterogeneity) [8-11]. Therefore, *personalized medicine* aims at delivering individualized therapies based on genetic and molecular profiles of individual patients, that may involve re-purposing of known drugs to different patients groups, hence helping ease the pharmaceutical industry bottlenecks related to the cost and time required to develop new drugs [11-12]. Methods for network data analytics and integration will be fundamental to these nascent areas, as full understanding can only come from holistically mining all available genetic, molecular and clinical data [11-12].

Holistic analyses of our interconnected world call for conceptual and methodological paradigm shifts. Rather than analyzing a single data source in isolation, such as aligning genetic sequences (that has already revolutionized our understanding of biology) [14], further insights will come from aligning all types of data within a single framework, that we term “the data alignment.” For example, all genetic and molecular interaction data about a cell can be integrated within the same computational framework, and we need to develop methods for aligning these “integrated cells” within a new paradigm of “the cell alignment.” Similarly, the world’s economic system includes networks of trade, financial exchanges and investments, which thus far have been studied individually [3,5], but a complete understanding of the origins of wealth, crises, and economic recoveries can only come from aligning and collectively analyzing all of these layers of networked economic and geo-political data. Likewise, climatic measurements are captured by various network types encoding the relationships between climatic elements across geographical regions (e.g., wind speed, atmospheric pressure and temperature) [13] and holistic, data-aligned analyses may help us understand this complex, dynamic system and better predict the effects of human caused alterations. Mathematical formalisms capable of capturing the intricacies of higher order organization of the data, along with the algorithms to compute and extract information from those formalisms need to be developed and applied [15]. Extending the framework of Benson *et al*. towards finding higher-order structures within these integrated and aligned data systems may be a way forward. Computational issues remain to be addressed, arising from large sizes, complexity, heterogeneity, noisiness, and different time and space scales of the data.

References

[1] M.R. Garey and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W. H. Freeman & Co. New York, NY, USA (1979).
[2] R. Benson *et al.*, *Science* X,X (2016).
[3] O.N. Yaveroglu *et al.*, *Scientific Reports* **4** (2014).
[4] K. Steinhaeuser and A. A. Tsonis, *Climate Dynamics* **42**, 5 (2014).
[5] P. Glasserman and H. P. Young, *Journal of Banking and Finance* **50** (2015).
[6] R. Sharan *et al.*, *Molecular Systems Biology* **3**, 1 (2007).
[7] A.L. Barabasi, *Science* **325**, 5939 (2009).
[8] A.L. Barabasi *et al.*, *Nature Reviews Genetics* **12**, 1 (2011).
[9] J. Menche *et al.*, *Science* **347**, 6224 (2015).
[10] M. Zitnik *et al.*, *Scientific Reports* **3** (2013).
[11] V. Gligorijevic *et al.*, *Proteomics* **16**, 15 (2016).
[12] S. M. Strittmatter, *Nature Medicine* **20**, 6 (2014).
[13] D. Davis *et al.*, *Bioinformatics* **31**, 10 (2015).
[14] J. Alfoldi and K. Lindblad-Toh, *Genome Research* **23**, 7 (2013).
[15] S. Boccaletti *et al.*, *Physics Reports* **544**, 1 (2014).

Acknowledgments

This work was supported by the European Research Council (ERC) Starting Independent Researcher Grant [278212], the National Science Foundation (NSF) Cyber-Enabled Discovery and Innovation (CDI) [OIA-1028394], the ARRS project [J1-5454] and the Serbian Ministry of Education and Science Project [III44006].

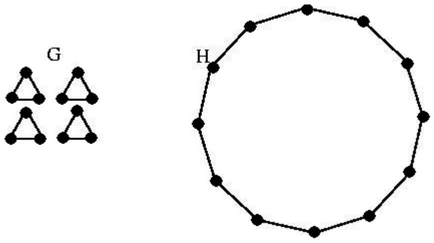

Fig. 2: An example of two networks, G and H, of exactly the same size (the same number of nodes and edges) and with each node in each of them having the same degree, but that are of very