Network analytics in the age of big data

How can we holistically mine big data?

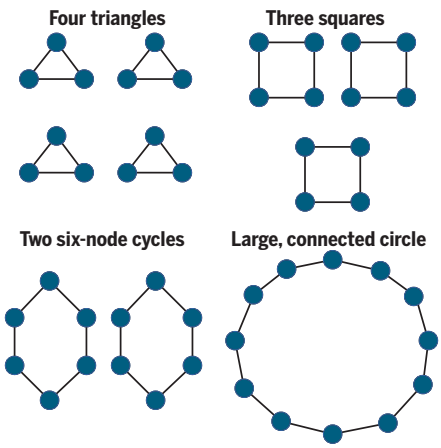
By **Nataša Pržulj** *and* **Noël Malod-Dognin**

We live in a complex world of interconnected entities. In all areas of human endeavor, from biology to medicine, economics, and climate science, we are flooded with large-scale data sets. These data sets describe intricate real-world systems from different and complementary viewpoints, with entities being modeled as nodes and their connections as edges, comprising large networks. These networked data are a new and rich source of domain-specific information, but that information is currently largely hidden within the complicated wiring patterns. Deciphering these patterns is paramount, because computational analyses of large networks are often intractable, so that many questions we ask about the world cannot be answered exactly, even with unlimited computer power and time (*1*). Hence, the only hope is to answer these questions approximately (that is, heuristically) and prove how far the approximate answer is from the exact, unknown one, in the worst case. On page 163 of this issue, Benson *et al.* (*2*) take an important step in that direction by providing a scalable heuristic framework for grouping entities based on their wiring patterns and using 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 (see the figure). Instead, Benson *et al.* use higher-order descriptors called graphlets (e.g., a triangle) that are based on small subnetworks obtained on a subset of nodes in the data that contain all interactions that appear in the data (*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 group together 20 neurons in the nematode worm neuronal network that are known to control a particular type of movement. 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.

Network structures

The four networks shown have exactly the same size (the same number of nodes and edges), and each node in each network has the same degree (the number of interactions with other nodes), but each network has a very different structure.



GRAPHIC: N. PRŽULJ AND N. MALOD-DOGNIN/ADAPTED BY V. ALTOUNIAN/*SCIENCE*

The importance of this result lies in its applicability to a broad range of networked data that we must understand to answer fundamental questions facing humanity 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 adopt various three-dimensional structures to carry out particular cellular functions. Molecular interactions are captured by different high-throughput biotechnologies and modeled 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 (*13*), leading to better understanding of gene functions (*6*) and molecular organization of the cell (*7*) and to improved therapeutics (*8–12*).

However, each network type provides limited information about the phenomenon under study. For example, a disease is rarely the consequence of a single mutated gene, or of a single broken molecular interaction. Rather, it is the product 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*). By 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 repurposing of known drugs to different patient groups, hence helping to 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*).

> ***"To mine the wiring patterns of networked data and uncover the functional organization, it is not enough to consider only simple descriptors...."***

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 (which has already revolutionized our understanding of biology) (*14*), further insights will come from aligning all types of data within a single framework—"the data alignment." For example, all genetic and molecular interaction data about a cell can be integrated into the same computational framework, and methods need to be developed for aligning

Department of Computer Science, University College London, London, UK. Email: natasa@cs.ucl.ac.uk

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 geopolitical 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) (*4*), and holistic, data-aligned analyses may help to explain 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, should be developed and applied (*15*). Extending the framework of Benson *et al.* to 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. ■

"Holistic analyses of our interconnected world call for conceptual and methodological paradigm shifts."

REFERENCES AND NOTES

1. M. R. Garey, D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, New York, 1979).
2. A. R. Benson *et al.*, *Science* **353**, 163 (2016).
3. O. N. Yaveroğlu *et al.*, *Sci. Rep.* **4**, 4547 (2014).
4. K. Steinhaeuser, A. A. Tsonis, *Clim. Dyn.* **42**, 1665 (2014).
5. P. Glasserman, H. P. Young, *J. Bank. Financ.* **50**, 383 (2015).
6. R. Sharan *et al.*, *Mol. Syst. Biol.* **3**, 1 (2007).
7. K. Mitra *et al.*, *Nat. Rev. Genet.* **14**, 719 (2013).
8. A.L. Barabási *et al.*, *Nat. Rev. Genet.* **12**, 1 (2011).
9. J. Menche *et al.*, *Science* **347**, 6224 (2015).
10. M. Žitnik *et al.*, *Sci. Rep.* **3**, 3202 (2013).
11. V. Gligorijevic *et al.*, *Proteomics* **16**, 741 (2016).
12. S. M. Strittmatter, *Nat. Med.* **20**, 590 (2014).
13. D. Davis *et al.*, *Bioinformatics* **31**, 1632 (2015).
14. J. Alfoldi, K. Lindblad-Toh, *Genome Res.* **23**, 1063 (2013).
15. S. Boccaletti *et al.*, *Phys. Rep.* **544**, 1 (2014).

ACKNOWLEDGMENTS

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

10.1126/science.aah3449

ENGINEERING

Solar-powering the Internet of Things

Photovoltaics can help to power remote sensors and controllers at the edge of the Internet

By **Richard Haight, Wilfried Haensch, Daniel Friedman**

The Internet connects billions of computational platforms of various sizes, from supercomputers to smart phones. However, the same types of data transmission can connect computational resources to much simpler sensors "at the edge of the net" that collect, analyze, and transmit data, as well as controllers that receive instructions. Devices deployed in the environment, homes and offices, and even our bodies would expand the number of connected devices to the trillions. This "Internet of Things" (IoT) underlies the vision of smart homes and buildings that could sense and transmit their status and respond appropriately (*1*), or track and report on the state of objects (vehicles, goods, or even animals) in the environment. However, the practical implementation of the IoT has been relatively slow, in part because all of these edge devices must draw electrical power from their local environment. We analyze the use of photovoltaics (PV) to power devices and help bring the IoT to fruition.

Wide-scale deployment of devices to remote or inaccessible areas while providing operational power in the absence of wires would require harvesting of available energy to ensure long-term operation. Indeed, one barrier to making an existing building "smart" is the cost of installing wired devices, so devices simply mounted to walls and powered by ambient indoor lighting are attractive for lowering installation costs. Whatever the specific form of energy used (alternatives to PV include kinetic, thermal, or wind energy), it must have sufficient energy density and reliability. The power requirements are dictated by the three components that process, sense, and communicate data. An excellent example of a system at the far edge of the net is the ultralow-power chip developed at the University of Michigan that incorporates much of the required functionality and energy-harvesting infrastructure (*2*).

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA. Email: rahaight@ibm.us.com

An ultralow-power processor that might, for example, be used in a smart phone (*3*) would consume ~85 μW of power when operating at a clock frequency of 1 MHz. Active power consumption in a processor is given by CV^2F, where C is the capacitance, V is the operating voltage of the processor, and F is the clock frequency. Reducing the operating voltage has the biggest benefit, and energy-efficient processors have been demonstrated to run with a supply voltage of <400 mV at 0.83 MHz (*4*). The time constant of the systems monitored is often long (milliseconds to minutes), so the required clock frequency can be well below 1 MHz. Intermittent operation—where the processor can be woken to take a measurement, store data, and later transmit it to a receiving station—would save appreciable power. Hence, ultralow-power operation of an edge device is both required and achievable.

Sensors are another area where low power is critical. As an example, consider an electrochemical sensor (*5*) used to monitor concentrations of toxic gases or other air-dispersed pollutants. A typical sensor might consist of a miniature electrochemical cell that draws 100 to 500 μA of current at 1 to 2 V. The total power consumed depends on whether it operates continuously (as for a carbon monoxide monitor) or periodically (an outdoor ozone monitor for air quality). Typical power requirements for most sensors can range from 100 μW to >1 mW, depending on what is being sensed.

Another critical component of an autonomously deployed edge device is its ability to transmit data to other proximal devices or a central location (hub) where the information can be analyzed. There are several communication modalities that could be used that operate near 1 V but differ in cost and distance range. A simple method would be to transmit the data optically by modulating an infrared light–emitting diode that operates at milliwatt power levels (but limited distance range). Vertical-cavity surface-emitting lasers are highly directional, providing longer range, but operate at several milliwatts. Yet another communication mode involves radio-frequency (RF) transmission being developed for ultralow-power operation in

Network analytics in the age of big data

Natasa Przulj and Noël Malod-Dognin

Science **353** (6295), 123-124.
DOI: 10.1126/science.aah3449