Network Science

A framework for finding commercial use cases with an example

Dan Herweg

May 2019

MADAS

Prepared for: Matteo Morini

Introduction

Network Science and offers an approach to problems involving large numbers of interacting nodes/agents. These techniques offer a rich new perspective on systems that have seen utility in academic settings and are increasingly exploited in commercial applications. This paper will provide a framework to find applications for the insights from Network Science and quantify the financial value of this information. The framework is described in the body and an initial exercise is included as an appendix.

So, while it is generally said that one should start with the problem and then select the appropriate tool, here we will take our hammer and go looking for nails. Maybe it is a fools's errand, but there is also evidence that application lags scientific advances by about 20 years [1]. Considering Watts and Strogatz' paper on small world networks (1998) and Barabasi's paper on scale free networks (1999), perhaps we are right in time.

Methodology

This framework has four steps:

1) In the first iteration, survey several major texts in Network Science for things that can be predicted analytically about nodes, links, networks and their evolution
2) Consider several businesses and the potential systems of relevance to each business, especially those with large numbers of interacting agents. Identify the relationships, actions, interactions, environment, entry and exit of these agents, and note whether they represent a directed, undirected, weighted, unweighted, multipartite, multilayered, multimode, or temporal graph. Here it will be simplified to nodes and links for space
3) Turn the list of "things we can say/predict about a network's nodes, links, and evolution" into things we can say about businesses and assess the availability of data
4) Quantify what each of these applications might mean to business decision makers in terms of costs, revenues, risks, etc and estimate this information's value to the node, a subset of nodes, or the network owner, as appropriate.

This is a process that should be repeated. Step 1 should recur as one learns more Network Science, as more Network Science is discovered, and as computational power increases. As new businesses are considered or new data becomes available, step 2 can be repeated. Step 3 might deserve to be revisited as new insights arise, and step 4 may need reevaluation as business conditions change. I have included here a first output of steps 1 and 2, and, due to the combinatorial size of an unconstrained search, one outline of how steps 3 and 4 could look.

Step 1 : What can we compute about networks?

As we move from nodes to networks we move towards less obvious metrics, more computationally-intensive techniques, more complexly reasoned interpretations. These are where the most value is expected to be added. The following is an outline of the initial list that will be used to generate analysis ideas, which should be periodically updated. Some metrics are expected to be largely trivial, but are listed here for completeness.

Node-Level Metrics

Many node characterizations are used to determine which nodes are important in specific ways. Does this node have many connections, connect many unconnected nodes, or are its connections dense? Stability looks at a nodes neighbors and patterns of liking/disliking and may reveal whether a relationship will persist in time.

> Local clustering coefficient, path lengths, degree, overlap, centrality, number of paths between nodes, bridgeness, balanced triads/pairwise stability [2]

Network-Level Metrics

At the network level, we first use metrics to describe the topology of the network.

> Distance, distribution of shortest paths, global clustering, average path length, degree distribution, degree correlation, motifs (causal, temporal), component size distribution, detection of communities, homophiliy, temporal entry and exit of nodes or links, assortativeness [3]

Then we can ask what this topology implies. These questions are contingent on having done the analysis on real data, so this is left as a reminder for future reviews of this document rather than something that can be addressed in this submission.

a. How could this network have evolved? How is it evolving now? Do we want to influence this? Can we characterize, for example, the superlinearity of preferential attachment?
b. What heuristics work best in this network for various tasks – spreading information, finding short paths, containing outbreaks, disrupting components
c. Is this structure robust or resilient? What would improve it? [4]

Step 2: What are some networks relevant to businesses?

To go about enumerating systems, I found it useful to first will divide them into internal and external systems. This may reflect the collection and governance of the data surrounding these systems and decisionmakers ability to influence them. I used large firms and governments to brainstorm here, since they might have better data and hire more Network Scientists as consultants.

External Networks

Customers, potential hires, suppliers, competitors, the public, power grids, animal life, geographical networks, public infrastructure

Internal Networks

Employees, departments, business units, assets, computer networks, private infrastructure

Step 3: What can we say about business networks?

What mathematical or predictive relationships can tell us

When there is a relationship between two or more variables, there are several potentially relevant things that can be said about the world. To begin, the relevant variables should be categorized according to a decision makers influence upon them: controllable and uncontrollable. Also of importance is the availability of the data, but for now the analysis will be unconstrained. The potential to get new insights may even be a sufficient incentive to collect new data.

Next, the relationships among the variables can be used for analysis. Any of the variables that are suspected or supposed to be causally related can be adjustable, constrained, or solved for. The results of this analysis can be valuable, as well as the process itself. Analysis may provide a base upon which to construct new tests, such as testing causality.

Additionally, probabilistic relationships can give ranges of predicted values and the chance of their realization, which enables risk and scenario modeling. If the relationships are not known to be causal, their causality can be tested or new variables can be sought to explain non-causal correlations. Many of these calculations can be done to characterize the state of a model or look forward to predict or influence the future state.

The permutations of solving for one variable while varying or constraining the others are numerous. In a decision-making context, factors such as maximization/minimization goals and the capacity to influence a variable will narrow the field of relevant analyses. In a firm-specific context, the relevant things a decision maker might want to know may be even more finite.

Step 4: What is the value of these techniques?

How to think about commercial viability

To be commercially viable, the methods will have to solve business problems. That is, they will have to be capable of helping business leaders make decisions which result in greater revenues or reduced costs. Since much analysis is already being done in commercial institutions, these methods will have to either differentiate themselves by providing answers to new questions, better answers to old questions, or be able to provide equivalent answers to old questions at a lower cost. It will also be helpful for the result to suggest a decision or intervention that can be made.

Effective, persuasive communication by specialists to non-specialist audiences is outside the scope of this paper, but it should be noted as an important prerequisite for increased adoption of these methods.

Here prediction alone is considered, as both a starting point and as the most familiar and quantifiable type of endeavor in business analysis. However, there are many other valuable outcomes from analysis, though less easily quantified, including explaining a phenomenon, illuminating uncertainties, and discovering new questions. [5]

Since network science involves assembling datasets that are wide reaching and often implicit, they may require new data collection or interpretation. This might mean that the true opportunity is to create systems that allow better tracking of relevant networks.

As a final consideration, large and connected networks are known for their ability to create novelty through the interaction of large amounts of diverse agents. Methods that strengthen this serendipitous exchange may also produce valuable innovations.

What follows is a first iteration of thinking about network science using this framework, with a hypothetical analysis that might result from it.

Step 1: What can be said about networks?

Examples of Network Analysis	General Description
Clustering coefficient	Describes how likely are a node's neighbors to be neighbors, and can be calculated globally as an average for all nodes or a ratio of closed triads
Path lengths	How far nodes are from each other, how far nodes are on average in a network, the distribution of these lengths
Degree	At the node level this is trivially the number of links, but the network distribution can give insight into network characteristics (randomness, small worldness, scale freeness), and evolution (growth, preferential attachment)
Overlap	Details how nodes have overlapping neighbors and the size of the set of their combined neighbors vs. their individual degree
Centrality	Determines the importance of a node or the distribution of a measure of importance in a network. Various flavors of centrality can determine importance with respect to whether nodes are "popular," intermediaries, or have many indirect links
Bridgeness	Determines if a node is likely to connect two large components
Balanced triads/pairwise stability	Determines if relationships are stable
Configuration model	This can be used to isolate network characteristics to determine in a simulation whether they cause an observed phenomena
Motifs (causal, temporal)	Patterns of links and interactions that are repeated in networks, often with a temporal aspect, and how frequent they are seen in a network
Homophily	Are nodes in a network more likely to be connecdted to nodes with similar features?
Community Detection	Which nodes can be grouped together coherently?
Entry/Exit of nodes/links	Obvious statistic that may have non-obvious impact on network structure (i.e. will it be scale-free)
Assortativeness	The correlation between the degree of a node and the degree of its connections

Step 2: What are some networks relevant to business

Potential Nodes
Buyers
Suppliers
Competitors
Innovators
Employees, current or potential
Recruiting potential new employees
Infrastructure nodes, capital or computer networks
Twitter followers
Citizens
Soldiers (Government)

Potential Links
Social (kin or friend networks, email communications)
Financial (money spent, number of transactions, client contacts, advertising)
Common features (type of firm, characteristics of consumer)
Collaborations (patents, academic papers, work teams)
Physical (geographical distance, infrastructure)

Example of Steps 3 & 4

Step 3: What can be said about business networks?

Theory

Robustness of a network is generally defined as how many nodes or links must be removed to destroy the giant component. This number can depend on how scale-free a network is. Scale-freeness can be characterized by the degree distribution and evolutionary process of node entry and how it is linked – it must exhibit growth and preferential attachment.

Business Case

Lets say you are Amazon, internalizing transportation costs and developing scale and know-how by building a new transportation network by building warehouses and buying trucks to go between them. This network delivers products across a geography that is prone to inclement weather that will close roads on occasion. This will either temporarily remove links between nodes or nodes themselves if the employees cannot come to operate the warehouse. The incremental growth of trucks and warehouses could be driven by robustness and resilience concerns. Preserving the giant component means making all deliveries on time by connecting a high percentage of nodes to move products to their destinations.

Analysis

Proposing new trucks or warehouses could be subjected to analysis of how added links and nodes affect the system and change the critical threshold of failures that are necessary to cause failures. Failure rates and acceptable service levels could be assumed to solve for the cheapest next step that maintains service, or the cost of the next step could be considered in the light of the benefits to service it would bring versus another option.

Step 4: What is the value?

Cost-Benefit Calculations

What savings from potential missed deliveries exist?
What is the value of the company's reputation for reliability in future revenue?
What infrastructure cost is recommended based on the analysis vs. alternatives?
Does the recommendation change the geographies that can be served and the associated revenues?
What is the impact of changing average path lengths on costs such as gas?
What does data gathering and analysis cost?

References

[1] S. W. J. G. Zoë Slote Morris, "Theansweris17years,whatisthe question: understanding time lags in translational researc," *Journal of the Royal Society of Medicine,* no. 104, pp. 510-520, 2011.

[2] S. K. Mohammed Zuhair Al-Taie, Python for Graph and Network Analysis, Cham: Springer Nature, 2017.

[3] M. O. Jackson, Social and Economic Networks, Princeton: Princeton University Press, 2008.

[4] A.-L. Barabasi, Network Science, Cambridge: Cambridge University Press, 2016.

[5] J. M. Epstein, "Why Model?," *Journal of Artificial Societies and Social Simulation,* vol. 11, no. 4 12, 2008.