JOURNAL OF INFORMATION TECHNOLOGY CASE AND APPLICATION RESEARCH
https://doi.org/10.1080/15228053.2023.2187995

EDITORIAL

Network analytics: an introduction and illustrative applications in health data science

ABSTRACT
Analytics researchers are widely using network analysis as a part of their methodology. In this review paper, we discuss different network concepts while summarizing some studies conducted using descriptive, predictive, and prescriptive analytics approaches. These applications illustrate the value of incorporating network properties of a phenomenon in better understanding the problem, prediction, and optimization of an outcome of interest, especially in the health domain.

Introduction

Network Analysis is a popular method for analyzing complex problems involving interactions among features or observations. While network analysis is not a new technique, it has recently gained momentum due to the availability of cheap computing as the algorithms to analyze large networks require large processing power. In addition, its suitability for analyzing large datasets involving underlying relationships or connectedness has made it one of the top choices for analytics researchers.

A network comprises nodes connected through well-defined edges. One major area that generates network-type data is social media, where relationships are explicitly embedded. In other words, the nodes make decisions to connect to other nodes. For instance, two friends on Facebook make a connection in the Facebook network, which is an explicit edge. However, there are other types of networks with implicit relationships that are defined using some underlying exchanges derived using some computation. Examples include product co-purchase network (Dhar et al., 2014), ingredient network (Teng et al., 2012), comorbidity network (Hidalgo et al., 2009; Kalgotra et al., 2017), text-based network (Celardo & Everett, 2020), brain parts network (Kalgotra & Sharda, 2018) and others.

In this paper, our focus is on discussing a variety of data science research emerging using network analysis. It is important to note that our focus is not specifically on social networks. We refer the interested reader to a review paper by Borgatti et al. (2009) in the *Science* journal in which the authors elaborated on the history of social network analysis, the theories emerging from the social network analysis, and the type of research questions studied in the past.

In this review paper, we start by discussing the representation of network data and common outputs from network analysis. Then, different types of network analytics research are discussed. In each type, we include some of our published papers as examples in addition to other papers. Finally, we conclude by describing the potential contributions of network analytics research.

Network data and analysis output

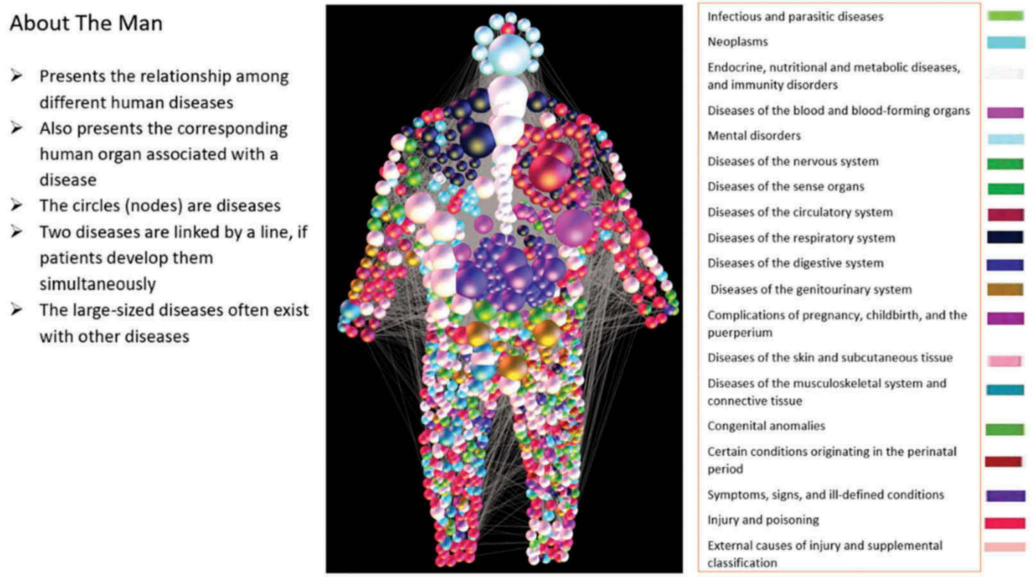
As noted, previously, there are two main elements in a network: nodes and edges. To connect the nodes, the edges are required to be defined. If the connections are binary, i.e., present or absent, the edges are unweighted. In contrast, if the strength of the ties between the nodes varies, the edges are weighted. Moreover, if there is a direction in a connection between one node to the other, the edges are directed, whereas the edges are undirected if there is no direction in the relationship between two nodes. Therefore, a clear definition of the nodes and edges is required in the network analysis. Consider an example of 100 friends forming an online social network. These 100 friends can form an unweighted, undirected network based on who knows whom. At the same time, the same 100 friends can form another weighted directed network based on the frequency/volume of conversations between specific friends. The latter network could be quite different from the first network; that simply describes who is connected to who. Thus, it entirely depends on the definition of the nodes and edges.

Unlike the above networks, where the relationship between nodes is explicitly defined by the nodes or by a user, in an implicit network, the network itself is derived from the underlying relationship between the elements. It requires an additional step to organize the data in the form of nodes and edges. For instance, in a co-occurrence network where edges are defined based on the presence of certain nodes in the same transactions, the nodes and edges are derived from a transactional dataset. The common metrics applied in the network research to derive the edges include Jaccard's index (Jaccard, 1901), Salton Cosine index (Salton & McGill, 1983), pointwise mutual information (Teng et al., 2012), relative risk, and correlation coefficient, among others. For example, Kalgotra et al. (2017) used Salton Cosine Index to define an edge in the comorbidity network based on the co-occurrence of diseases in the patients. The set of patients with multiple diseases during the hospital visits made the transactional set from where the edges between the nodes were derived using a similarity index. Common representations of the edges in a network include an adjacency matrix, an adjacency list, and an edge list. The network dataset is typically stored in a graph database.

The most common outputs obtained from the network analysis are the node centralities such as degree, closeness, betweenness, and eigenvector. Other measures at the node level may include clustering coefficient and page rank, among others. At the structural level (macro), the metrics such as distribution of centralities, average path length, and density result in a typical network analysis. The micro (node) and macro (structure) level measures are used to understand the properties of the network. For instance, Watts and Strogatz (1998) used cluster coefficient and path length to derive the small-world phenomenon, whereas Barabási and Albert (1999) used the distribution of the degree centrality to describe the scale-free property of the network.

In addition to the mathematical computations, the common output from a network analysis is a visualization. Since it is difficult to make sense of the visualization of a network consisting of a large number of nodes and edges, visualization researchers have developed different layouts to construct a view of the network. Common layouts include Fruchterman-Reingold (Fruchterman & Reingold, 1991), Yifan Hu Multilevel (Hu, 2005), etc. Although a layout simplifies the structure of the network by reorganizing the nodes in a visual space, it does not provide precise information about the network, especially when the number of nodes is large. A network visual is of little value if it is not annotated properly or presented creatively. Therefore, the onus is on the researchers to present the visual by annotating it or creatively organizing it. A good example of the annotation of a network can be found in Hidalgo et al. (2009). On the other hand, Figure 1 presents a network of diseases developed by the authors of this paper. In this undirected comorbidity network, the nodes are diseases. Two diseases are connected if these co-occur in the patients. The strength of a connection is computed using Salton Cosine Index. To provide meaning to the network, we organized the network in the shape of a human body and placed the diseases at the corresponding organ system. The size of a node is based on the degree centrality. It is easy to interpret from the visual that mental disorders and heart disorders have the highest degree, which would have been difficult to infer using a typical network layout. The same network was presented in Kalgotra et al. (2017) with the Fruchterman-Reingold layout.

The original method of presenting the network is a graph matrix, which is a mathematical representation of the network. With the advent of graphical user interfaces, the two-dimensional visualizations of large networks became popular with software such as UCINET (Borgatti et al., 2002), Gephi (Bastian et al., 2009), etc. We expect the next step in the visual network analysis to be the analysis through virtual and augmented reality, which is more immersive and will likely increase the adoption of the network method further across the disciplines (See Figure 2). It seems to be the natural evolution of network analysis. Some researchers and companies are already exploring this idea of network analysis in virtual reality such as VRNetzer (Pirch et al., 2021).



About The Man

- Presents the relationship among different human diseases
- Also presents the corresponding human organ associated with a disease
- The circles (nodes) are diseases
- Two diseases are linked by a line, if patients develop them simultaneously
- The large-sized diseases often exist with other diseases

Infectious and parasitic diseases
Neoplasms
Endocrine, nutritional and metabolic diseases, and immunity disorders
Diseases of the blood and blood-forming organs
Mental disorders
Diseases of the nervous system
Diseases of the sense organs
Diseases of the circulatory system
Diseases of the respiratory system
Diseases of the digestive system
Diseases of the genitourinary system
Complications of pregnancy, childbirth, and the puerperium
Diseases of the skin and subcutaneous tissue
Diseases of the musculoskeletal system and connective tissue
Congenital anomalies
Certain conditions originating in the perinatal period
Symptoms, signs, and ill-defined conditions
Injury and poisoning
External causes of injury and supplemental classification

Figure 1. Comorbidity network.

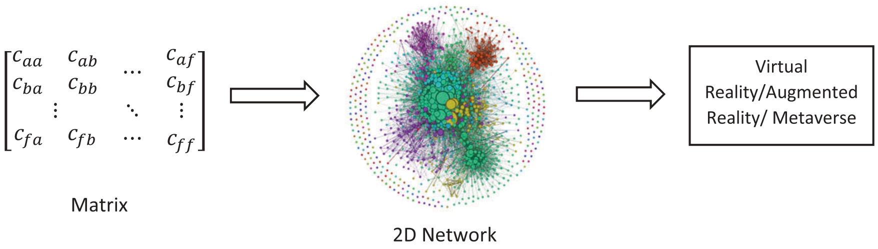

$\begin{bmatrix} c_{aa} & c_{ab} & \cdots & c_{af} \\ c_{ba} & c_{bb} & & c_{bf} \\ \vdots & & \ddots & \vdots \\ c_{fa} & c_{fb} & \cdots & c_{ff} \end{bmatrix}$
Matrix
2D Network
Virtual Reality/Augmented Reality/ Metaverse

Figure 2. Evolution of network output.

Types of network analytics research

Most analytics studies use the above-discussed network measures and outputs in one or another forms. However, researchers have used these innovatively to perform descriptive, predictive, and prescriptive analytics research. In this section, we discuss some examples of each type of analytics.

Descriptive network analytics

In descriptive analytics, the broad purpose is to find interesting patterns in a large historical sample (representing the population) and develop novel questions or observations about what is being observed. One interesting descriptive study was performed more than two decades ago by Albert et al. (1999) in which the authors demonstrated the network of websites on the World Wide Web (WWW). The purpose was to find the diameter of the WWW network. The authors found that any two randomly selected web pages were only 19 connections away from each other, indicating the small world phenomenon in the WWW. The study further argued that with the 1000% increase in the number of web pages, the average diameter will only increase by 2 units. This is an interesting descriptive analytics study that found WWW to be a small network despite billions of nodes in it. Such an analysis of the current WWW can yield interesting findings.

Another application of descriptive analytics is our study on health disparities by race (Kalgotra, Sharda, & Croff, 2020), in which we identified the comorbidity differences in seven different races through comorbidity networks developed from patient-level data. The seven races were White, African American, Asian, Hispanic, Native American, Bi- or Multi-racial, and Pacific Islander. By using more than 18 million patients' medical records, one comorbidity network was developed for each race based on the multiple diseases co-occurring in the patients. To define an edge, Salton Cosine Index was used. The comorbidity networks of races were compared to find unique and common comorbidities across the races. For instance, the relationship between infectious and parasitic disorders with respiratory, circulatory, and genitourinary system disorders was stronger among African Americans than others. On the other hand, the connection of mental disorders with respiratory, musculoskeletal system, and connective tissue disorders was more prevalent in Whites than in other races.

Likewise, we conducted a similar network analysis to find comorbidity differences between men and women (Kalgotra et al., 2017). A comorbidity network for men and another one for women was created. The women's comorbidity network was denser than the men's network. Specifically, mental disorders had a higher number of connections with other diseases in women's multimorbidity network than in the men's network. On the other hand, the connection of chronic heart disorders with other disorders was stronger in men than women.

Another interesting example of descriptive analytics is an ingredients network created by Teng et al. (2012) in which the authors used the co-occurrence of multiple ingredients in the same recipe to create the network and identify the clusters of ingredients to suggest new recipes. Similar other descriptive analytics applications are available that utilize network analysis as a core methodology to model interactions in large datasets from the business and other scientific domains.

Predictive network analytics

Since the main criticism of network analysis is its descriptive nature as discussed by Borgatti et al. (2009), many researchers have attempted to create predictive analytics applications. More recently, we are seeing network analytics techniques used in conjunction with other approaches, such as deep learning or natural language processing.

There are two potential types of predictive analytics using the network method. First, the attributes of the nodes or edges of a network are predicted. In this case, the attributes are an internal part of the network. One such example is a study by Dhar et al. (2014) in which the authors predicted the sales of books in a book co-purchase network created using a recommendation system on the Amazon website. In this example, books are the nodes and sales for the books were predicted based on the position of the nodes in the network. Another common type of research in the first category of predictive analytics is the link prediction problem in which the likelihood of the formation of new edges is estimated. The primary objective of link prediction studies is to model the evolution of a network. Lü and Zhou (2011) highlighted different methods and applications of link prediction in their survey.

Second and more recent, other external outcomes are predicted using novel features generated using network methodology as predictors. In this case, network analysis is a crucial part of the bigger methodology. One such study is by Kalgotra and Sharda (2021), in which network analysis has been used to predict an outcome exogenous to the network. Specifically, the comorbidity network was used to predict hospital length of stay (LOS). In this paper, the authors used electronic health records of more than 24.7 million patients across 662 US hospitals over 16 years (2000–2015). The authors used a two-step approach to create the machine learning model – creating comorbidity networks in the first step and then creating machine learning models for predicting LOS in the second step.

First, an independent sample of about three million patients was used to create comorbidity networks in which the diseases were the nodes, and two diseases were connected if they appeared in a patient during the same hospital visit. The comorbidity networks were used to create new variables for the remaining patients who were not part of the network analysis. To understand the new features generated using comorbidity networks, consider a patient who visits the hospital with a complaint of a hypothetical disease A. The only disease-related information available at the time of admission is disease A. In our application case, the network was searched for disease A, and the top five connected diseases were identified. These diseases were labeled as probable diseases as these were likely to be diagnosed during the hospital stay. In addition, a patient may have a history of diseases in the system, termed historical diseases. Together, the probable and historical diseases were called latent comorbidities. The new construct of latent comorbidities was then used in modeling and predicting LOS at the time of admission. The predictive models for LOS were created with patient demographics, the known diseases at the time of admission, and latent comorbidities as the independent variables. The Long-Short-Term Memory (LSTM) models were created without and with the latent comorbidities to compute the explanatory and predictive power added by the proposed variables. In terms of variance explained, the new construct added 3.6%, and in terms of mean absolute percent error (MAPE), the latent variable improved the MAPE by 1.9%. Although the numbers seem low, these are equivalent to the improvement in the forecast by $882.8 million. Therefore, the gain is practically significant.

As evident from the examples above, networks can be used for predictive modeling to gain additional predictive power. More such network-driven methods are required to predict outcomes that are endogenous or exogenous to the networks.

Prescriptive network analytics

The third type of analytics is prescriptive analytics in which the object is to find the optimal solution. Network analysis is a popular research area for prescriptive analytics. The majority of the prescriptive analytics research applying network analysis involves efficient processing and traversing of the network. Examples include creating prescriptive models to identify cliques in the network, as discussed by Miao and Balasundaram (2017), to traverse the network and identify the shortest path between nodes (Selim & Zhan, 2016), etc. Several Operations Research and Computer Science researchers are focusing on the topics of efficiently computing different network metrics. See Balakrishnan (2019) for such models. More prescriptive analytics applications in different domains are required to be explored using network analytics.

Conclusions

Network science has been used as a *theory* to understand an emergent phenomenon and as a *methodology* to model relationships. Although our focus is not on theory-driven network analysis, it is worthwhile to mention some important concepts and theories derived from network analysis. Some of the well-known concepts include random graphs (Erdős & Rényi, 1959), scale-free networks (Barabási & Albert, 1999), strength of weak ties (Granovetter, 1973), power law distribution of WWW (Adamic & Huberman, 2000), small world phenomenon (Watts & Strogatz, 1998), Benford's law in online social network (Golbeck, 2015), role of cliques (Provan & Sebastian, 1998), information Q1 diffusion (Bakshy et al., 2012), preferential attachment (Newman, 2001), network flow (Borgatti, 2005) and community detection (Reichardt & Bornholdt, 2006), among others. In addition, studies have been designed to understand the validity or the structure of the network (Kalgotra, Sharda, & Luse, 2020).

In data-driven research, the purpose is to discover new theories and patterns. Therefore, it is important to generalize the novel relationships between concepts involving network analysis. In network analytics studies, the contextual and methodological contributions are more apparent. In the papers with contextual contributions, novel networks are created. In other words, a problem is studied through a novel network lens. On the other hand, in the papers with methodological contributions, the network analysis is a crucial part of a bigger methodological process and thus, contributes to the method of the study. Subsequently, it is important to generalize the methodology so that it can be applied in different settings and different problem domains.

In this paper, we attempted to review different types of analytics research conducted using network analysis. In each type of analytics, several papers across the disciplines were discussed. In addition, the relevant concepts and references were listed throughout the paper. Therefore, our paper can be used as a guide by researchers and educators interested in learning and applying network methodology.

References

Adamic, L. A., & Huberman, B. A. (2000). Power-law distribution of the world wide web. Q2 *Science*, *287*(5461), 2115. https://doi.org/10.1126/science.287.5461.2115a

Albert, R., Jeong, H., & Barabási, A. L. (1999). Diameter of the world-wide web. *Nature*, *401* (6749), 130–131. https://doi.org/10.1038/43601

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012, April). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web* (pp. 519–528). Q3

Balakrishnan, V. K. (2019). *Network optimization*. Chapman and Hall/CRC.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286* (5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Bastian, M., Heymann, S., & Jacomy, M. (2009, March). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, *3*(1), 361–362. https://doi.org/10.1609/icwsm.v3i1.13937

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71. https://doi.org/10.1016/j.socnet.2004.11.008

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). Ucinet for windows: Software for social network analysis. *Harvard, MA: Analytic Technologies*, *6*, 12–15. Q4

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892–895. https://doi.org/10.1126/science.1165821

Celardo, L., & Everett, M. G. (2020). Network text analysis: A two-way classification approach. *International Journal of Information Management*, *51*, 102009. https://doi.org/10.1016/j.ijinfomgt.2019.09.005

Dhar, V., Geva, T., Oestreicher-Singer, G., & Sundararajan, A. (2014). Prediction in economic networks. *Information Systems Research*, *25*(2), 264–284. https://doi.org/10.1287/isre.2013.0510

Erdős, P., & Rényi, A. (1959). On random graphs. I. *Publicationes Mathematicate*, *6*(3–4), 290–297. https://doi.org/10.5486/PMD.1959.6.3-4.12

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force□directed placement. *Software: Practice & Experience*, *21*(11), 1129–1164. https://doi.org/10.1002/spe.4380211102

Golbeck, J. (2015). Benford's law applies to online social networks. *Plos One*, *10*(8), e0135169. https://doi.org/10.1371/journal.pone.0135169

Granovetter, M. S. (1973). The strength of weak ties. *The American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469

Hidalgo, C. A., Blumm, N., Barabási, A. L., & Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, *5*(4), e1000353. https://doi.org/10.1371/journal.pcbi.1000353

Hu, Y. (2005). Efficient, high-quality force-directed graph drawing. *Mathematica Journal*, *10* (1), 37–71.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat*, *37*, 241–272. Q5

Kalgotra, P., & Sharda, R. (2018). BIARAM: A process for analyzing correlated brain regions using association rule mining. *Computer Methods and Programs in Biomedicine*, *162*, 99–108. https://doi.org/10.1016/j.cmpb.2018.05.001

Kalgotra, P., & Sharda, R. (2021). When will I get out of the hospital? Modeling length of stay using comorbidity networks. *Journal of Management Information Systems*, *38*(4), 1150–1184. https://doi.org/10.1080/07421222.2021.1990618

Kalgotra, P., Sharda, R., & Croff, J. M. (2017). Examining health disparities by gender: A multimorbidity network analysis of electronic medical record. *International Journal of Medical Informatics*, *108*, 22–28. https://doi.org/10.1016/j.ijmedinf.2017.09.014

Kalgotra, P., Sharda, R., & Croff, J. M. (2020). Examining multimorbidity differences across racial groups: A network analysis of electronic medical records. *Scientific Reports*, *10*(1), 1–9. https://doi.org/10.1038/s41598-020-70470-8

Kalgotra, P., Sharda, R., & Luse, A. (2020). Which similarity measure to use in network analysis: Impact of sample size on phi correlation coefficient and Ochiai index. *International Journal of Information Management*, *55*, 102229. https://doi.org/10.1016/j.ijinfomgt.2020.102229

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications*, *390*(6), 1150–1170. https://doi.org/10.1016/j.physa.2010.11.027

Miao, Z., & Balasundaram, B. (2017). Approaches for finding cohesive subgroups in large□scale social networks via maximum k ‑ plex detection. *Networks*, *69*(4), 388–407. https://doi.org/10.1002/net.21745

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, *64*(2), 025102. https://doi.org/10.1103/PhysRevE.64.025102

Pirch, S., Müller, F., Iofinova, E., Pazmandi, J., Hütter, C. V., Chiettini, M., Menche, J. . . . Menche, J. (2021). The VRNetzer platform enables interactive network analysis in virtual reality. *Nature Communications*, *12*(1), 1–14. https://doi.org/10.1038/s41467-021-22570-w Q6

Provan, K. G., & Sebastian, J. G. (1998). Networks within networks: Service link overlap, organizational cliques, and network effectiveness. *Academy of Management Journal*, *41*(4), 453–463. https://doi.org/10.2307/257084

Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, *74*(1), 016110. https://doi.org/10.1103/PhysRevE.74.016110

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. Mcgraw-Hill.

Selim, H., & Zhan, J. (2016). Towards shortest path identification on large networks. *Journal of Big Data*, *3*(1), 1–18. https://doi.org/10.1186/s40537-016-0042-7

Teng, C. Y., Lin, Y. R., & Adamic, L. A. (2012, June). Recipe recommendation using ingredient networks. In *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 298–307). Q7

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Pankush Kalgotra
Auburn University, Raymond J. Harbert College of Business, Auburn, AL, USA
✉ pzk0031@auburn.edu

Ramesh Sharda
Oklahoma State University, Stillwater, OK, USA