

GRAPH-THEORETIC TOOLS FOR DYNAMIC NETWORK ANALYSIS

HÉLÈNE BARCELO* AND REINHARD LAUBENBACHER

ABSTRACT. Graph-theoretic tools have long been used in social network analysis. Typically, graphs are used to represent some type of connection between network members. Often, the network is changing over time and graph representation is an aggregate of a dynamic network trajectory. This paper introduces new graph-theoretic tools that capture information about network dynamics. Also, rather than representing a view of the entire network, the methods proposed here provide measures to study the dynamic behavior of individual network members. Two examples of applications are given, one to the analysis of social behavior of pre-school children, and the other to an epidemiological contact network.

1. INTRODUCTION

Graphs (with vertices and edges) have been extensively used in the social sciences, computer sciences and several other disciplines, to graphically represent various type of networks. In [15] Newman provides an excellent survey of the variety of techniques and models developed in recent years that help to understand or predict the behavior of such systems, while Wasserman and Faust ([16]) give an exhaustive description of the methods used in social network analysis. Data representation in social network systems can be encoded using graph theoretic notation. The nodes $\{n_j\}$ of the graph represent the set of actors, and an edge (arc) between two actors $\{n_i, n_j\}$ indicates a connection of some sort with each other.

More generally, in the one-mode network context, where structural variables are measured, one obtains a single graph representing the entire network ([16]). Various measures, wether at the individual or group level, such as closeness and centrality, continuing flow, prestige, and so on, are then extracted from the graph as a method of characterizing its actors. This method works well with limited observations but with more extensive observations, interpretation difficulties increase dramatically. Translating multiple observations into one graph, either using aggregation or weighted graphs, produces an unsatisfactory amalgam. Moreover, extracting valid information about an individual from such a graph is difficult. In particular, information gained from time course observations will likely be lost when aggregated into a single graph.

We propose here an approach that represents information gained from multiple observations of a dynamically changing network with the goals of 1) providing information about individual actors in the network and 2) measuring the dynamic evolution of the network. The mathematical theory that underlies this approach can be found in [3] and [4]. Our approach is to construct for each actor in the network a sequence of graphs that capture the network neighborhood of the actor at increasing levels of "persistence" over time. We then assign several numerical measures to this sequence of graphs which capture the breadth (extent of the network neighborhood) as well as the depth (persistence over time) of the actor's involvement in network evolution. The different measures can be aggregated into a single measure for each actor in several ways, reflecting particular aspects of its place in the network. The measure can then be used as the basis of clustering algorithms as well as global measures of the network, such as power law distributions. The initial motivation for our approach comes from Q-analysis, a method introduced by R. Atkin [1, 2] and further developed by

Date: January 19, 2006.

* Research supported by the NSA, MSPF-04G-110.

J. Johnson [12]. Q-analysis was developed further into a type of combinatorial homotopy theory, called A -theory [3, 4, 5], for which it represents a 0-dimensional invariant of a graph. We believe that it would be very profitable to use the higher-dimensional invariants of A -theory in social network analysis.

The paper is structured as follows. In Section 2 we introduce the mathematical basis for the method, and in Section 3 we present two examples of applications, one to the analysis of social behavior skills of pre-school children, and the other to an epidemiological contact network. Finally, in Section 4 we briefly describe a software package available for download as well as a web application, which implements this approach. The code is optimized to be able to handle networks with as many as a million nodes.

2. THE METHOD

We develop our method in a very general setting that applies to a variety of interaction networks, for instance most agent-based simulations. We assume that we are given a set $X = \{x_1, \dots, x_n\}$ of network nodes. In order to capture the interaction dynamics of the network, we assume that at a given time the state of the network is described by an $n \times n$ -matrix with binary entries. A 1 in entry (i, j) indicates that node x_i interacts with node x_j at that time, with a 0 indicating lack of interaction. The network evolves in discrete time, so that a trajectory of the network in time is given by a sequence of binary matrices.

Definition 1. A trajectory of an interaction network $\mathcal{X} = (X, M)$ is given by a set $X = \{x_1, \dots, x_n\}$ of n actors and a set $M = \{M_1, M_2, \dots, M_m\}$ of m matrices. As explained above, an (i, j) -entry of matrix M_r indicates that actor x_i interacts with actor x_j at time step r .

Note that in this definition the matrices M_i are not required to be symmetric, so that the notion of *interaction* is not necessarily symmetric.

We describe an example of an interaction network to which we will apply our analysis method in Section 3.

Example. Let X be a group of people, that interact with each other over time. The collection M of matrices could represent a time course of observations of who is interacting with whom. To be specific, suppose we are observing five children playing with each other over the course of four days. The observations are given in Table 1, and represent an interaction network with the five children $X = \{C_1, \dots, C_5\}$ as actors.

		Playmates				
		1	2	3	4	5
Child 1	Day 1	●	✓	✓	✓	●
	2	●	✓	✓	●	●
	3	●	●	✓	✓	●
	4	●	●	●	✓	●
Child 2	1	✓	●	✓	●	●
	2	●	●	✓	✓	●
	3	✓	●	✓	●	●
	4	●	●	●	●	●
Child 3	1	●	●	●	●	●
	2	✓	✓	●	●	●
	3	✓	✓	●	●	●
	4	✓	✓	●	●	●
Child 4	1	✓	✓	●	●	✓
	2	●	●	●	●	●
	3	●	●	●	●	●
	4	✓	●	✓	●	●
Child 5	1	●	●	●	✓	●
	2	●	●	●	●	●
	3	●	●	●	●	●
	4	●	●	●	●	●

FIGURE 1. Example of a network of playmates

The M_i are given by the collection of i th rows in the table. For example, $M_1(C_1) = \{C_2, C_3, C_4\}$, $M_1(C_2) = \{C_1, C_3\}$, etc. Therefore, M_1 is represented by the matrix

$$M_1 = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

where the 1's in the j – th row of M_1 represent the children who were playing with the *target* child j on day 1.

Our goal is to capture qualitative properties of the interactions observed in a trajectory by computing measures for the interaction "topology" of the entire time course. Rather than aggregating all observed interactions in a single graph we represent them by a sequence of graphs that capture the interaction "neighborhood" of each actor. In Definition 2 below we associate a sequence of graphs to each actor in an interaction network. We first illustrate this construction by way of the example above.

Example. Consider the observations of actor C_1 , for instance. For each $q = 1, \dots, 4$ we construct the graph G_1^q . Its vertices correspond to those children that were playing with child C_1 at least q times. There is an (undirected) edge between two vertices (children) C_i and C_j if there are at least q common observations. In this manner we obtain the sequence of graphs given in Figure 2. For example, in G_1^1 there is an edge between C_2 and C_4 since both children were playing with C_1 on Day 1, but there is no such edge in G_1^2 . This is because, while both played with C_1 on at least two days, they were not the same **two** days. Lastly, observe that for $q = 4$ all the graphs are empty and thus are not represented in the figure below.

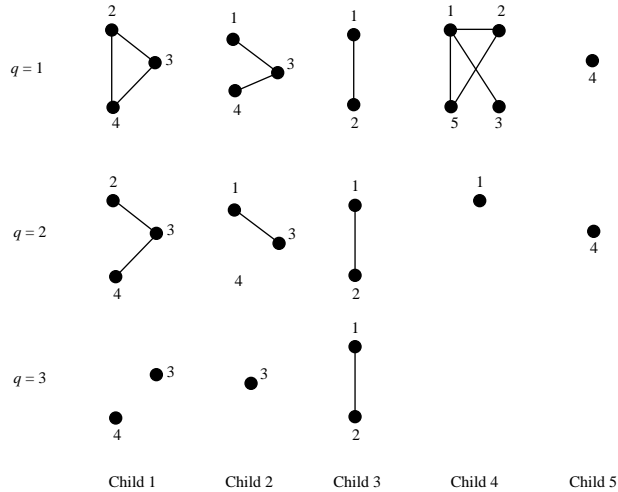


FIGURE 2. Sequence of graphs for each child.

We now formalize the construction in this example.

Definition 2. To each actor x_i we associate a sequence of m graphs

$$\mathcal{G}_i = \{G_i^1, G_i^2, \dots, G_i^m\},$$

described as follows. For $q = 1, \dots, m$ the vertices of G_i^q are given by those x_j for which $m_{ij}^k = 1$ (the (i, j) entry of the k^{th} matrices) for at least q values of k . That is, x_j was observed interacting

with x_i at least q times. Let x_j and x_h be two vertices (in G_i^q), let J be the set of such indices k for x_j , and let H be the corresponding set for x_h . Then there is an edge between x_j and x_h if the cardinality of the intersection $J \cap H$ is at least q .

As explained in the introduction, the graphs defined in this way differ significantly from the standard interaction graphs. Firstly, we do not aggregate the interactions over time. From the set of graphs $\{\mathcal{G}_i : i = 1, \dots, n\}$ we can reconstruct most of the time course of observations. For each individual actor x_i the sequence of graphs \mathcal{G}_i represents the network from this actor's point of view. The changes in the graphs, as q increases, are a measure of the stability and cohesion over time of the network in which the actor is embedded. We now describe some measures on the \mathcal{G}_i that can be used to compare actors and correlate graph structures to other available information about the actors.

Definition 3. We first define five real-valued vectors of length m (the number of observations) associated to each actor in an interaction network.

- (1) $V_i = (v_1^i, \dots, v_m^i)$, where v_q^i is equal to $|V(G_i^q)|$, the number of vertices in G_i^q .
- (2) $E_i = (e_1^i, \dots, e_m^i)$, where e_q^i is equal to $|E(G_i^q)|$, the number of edges in G_i^q .
- (3) $D_i = (d_1^i, \dots, d_m^i)$ be the vector with d_q^i equal to the density of the graph G_i^q . Recall that the density $d(G)$ of a graph G [include ref] is equal to

$$d(G) = \frac{2 \cdot |E(G)|}{|V(G)|(|V(G) - 1|)}.$$

- (4) $C_i = (c_1^i, \dots, c_m^i)$, where c_q^i is the number of connected components of G_i^q .
- (5) $R_i = (r_1^i, \dots, r_m^i)$, where r_q^i is the diameter of G_i^q , that is, the maximum of the lengths of minimal paths connecting any two vertices in G_i^q .

The first two vectors capture the changing size of the network neighborhood of actor x_i , as q increases. The other measures capture connectivity properties of the neighborhood. While an actor's neighborhood might be quite large, it might also be quite "shallow," in the sense that its connectivity to its neighbors might be very weak.

Using these vectors one can assign a variety of numerical measures to the actors in the network. For example, one can consider the Euclidean norm of the vectors. One can also associate several notions of distance between actors by using the Euclidean distance between corresponding vectors. More importantly, using data fittings techniques on each of the described vectors, one obtain various tools for comparing individuals' social networks. These are implemented in a software package available from the authors, and described in detail in [14].

3. APPLICATIONS

In this section, we briefly describe two applications of this approach, as an illustration of its utility and an indication of its generality. The first is to the analysis of data gathered in a study of peer-peer relationships among pre-school children [11].

Children's peer relationships are complex and multifaceted, and no single index has been able to capture this complexity. Furthermore, extant research has focused predominantly on static peer influences by examining peer relationships as they exist at a single point in time. Martin et al. [9, 10, 13] are exploring the dynamic shifts in the effects of peer influence on children's development and the formation of relationships over time. The goal of their study is to develop new ways of quantifying the dynamic properties of children's playgroup formation. Data collected as part of this study include periodic observations of several aspects of each child's context, including interactions with teachers and playmates. Observations extend over a period of 7-8 months during the school year. This preliminary study consisted in analyzing data collected on 97 children, with age ranging from 37 to 64 months, and attending three of the preschool classes at Arizona State University.

To each child one associates a series of 46 graphs, 46 being the number of days of observations in a given semester. Denote by G_q^i the q – *th* graph associated to the i – *th* target child, where q is an integer between 1 and 46 and i varies from 1 to 97. To simplify matters we build the graph G_q^X , that is the q – *th* graph associated to the target child X . The vertices of this graph correspond to the peer(s) with whom X interacted on at least q days during the semester. Thus, a child who interacts with Peer X on (exactly) 2 days would have a vertex in only two of the graphs G_q^X , namely for $q = 1$ and $q = 2$. In contrast, a child who interacts with peer X on (exactly) 4 days would have a vertex in 4 of the graphs G_q^X , namely for $q = 1, q = 2, q = 3,$ and $q = 4$. In this way the method captures the frequency of contact or exposure a target child has with a particular peer(s). There is an edge between two vertices Y and Z of a graph G_q^X , if at least q of the days for which Y and Z were interacting with X were the same.

In this manner we obtain a sequence of embedded graphs:

$$G_1^X \supseteq G_2^X \supseteq \dots \supseteq G_{46}^X.$$

That is, the second graph is a subgraph of the first one, the third a subgraph of the second and so on. Thus, once one of the graphs in the sequence is empty all the following ones are as well. A graph G_q^X is empty if there were no peers seen playing with the target child X on at least q days during the semester. The last integer q for which the graph G_q^X is *not* empty is called the *Qmax* measure for that child.

Note here that for each target child X , the same information could be recorded in a single *weighted* graph. Namely, the *weight* of an edge YZ would be the number of common days for which Y and Z were both playing with the target child X . While the information is the same, in the case of a single weighted graph it is not as easy to visualize the evolution of the playmates’ network of a given target child.

The preliminary analysis done by Hanish et al. ([11]) focused on the *qmax* measure as well as on the number of vertices in the sequences of graphs $G_1^X, G_2^X, \dots, G_{46}^X$ for all 97 children. Hanish et al. define two measures: *Qmax* and *Vertex scores*, the number of vertices in a graph. *Qmax* provides a measure of the *depth* of a target child’s relationship with a particular peer(s), namely the extent to which the target child interacts with a particular peer repeatedly over time. *Vertex scores*, represent the extensiveness or *breath* of a target child’s social network, namely the number of peers who are members of the child’s social neighborhood. Moreover, *vertex scores* can be calculated at each value of q , thus making it possible to estimate the breath of a child’s social network at different levels. The resulting graphs can be viewed at <http://nelligan.la.asu.edu/social/cgi-bin/graph.cgi>

We report here some the results of [11]. Older preschoolers had greater depth and breath (higher values of the *Qmax* measure, and higher vertex scores) in their relationships than younger ones. No significant differences between boys and girls were found with respect to the depth and breath measures. Children with high depth had significantly greater breath in their peer relationships than did children with medium and low depth.

Children with relatively short-term peer relationships (i.e., low depth) also have breath scores that rapidly drop over time, and form a more homogeneous group than the children with medium or high depth who appear to be more heterogeneous groups.

Findings also indicate that greater breadth was associated with more adaptive social functioning, with these findings strongest at greater levels of depth. Breadth was also positively correlated with pro-social behavior and effortful control, while negatively correlated with social withdrawal and peer victimization. Lastly, among children in the low depth group, these correlations were not significant.

As a second application we show how to apply our method to the analysis of contact data from a simulated epidemiological network. Details will appear elsewhere [14]. We use a data set generated by the interaction-based simulation program *EpiSims* [8], which was provided by S. Eubank of

the Simulation Science Laboratory at the Virginia Bioinformatics Institute. *EpiSims* simulates the spread of a pathogen in an urban area, as people move around the city and come in contact with each other. The data set we discuss here was generated by a simulation of the city of Portland, OR, with approximately 1.5 million inhabitants. The simulation generates a complete synthetic population, using demographic data, activity surveys, and a detailed implementation of the city's infrastructure and layout. From the simulation one can extract detailed information about people who are in the same location at the same time on a second-to-second basis. This allows the assembly of a detailed 24-hour trajectory, with a user-chosen resolution. Using an epidemiological model one can then attempt to answer questions such as which inhabitants contribute the most to the spread of the pathogen. Such information could be used to develop vaccination or quarantine strategies.

Each observation is given by a bipartite graph with approx. 3 million vertices. From a 24-hour *EpiSims* run we generated a trajectory of this interaction network by aggregating the data into observations that are 1 minute apart. Hence, the length m of the trajectory is 1440, with the number of actors $n = 1,500,000$. As described in Section 2, we extracted the sequence of graphs describing the interaction neighborhood of each of the 1.5 million inhabitants. We have computed some of the measures described in Section 2, which can then be used to cluster the population using an appropriate clustering algorithm. Each node is assigned a likelihood score using multivariate order statistics, measuring the expected structure of its neighborhood in terms of the number and size of components. Figure 3 shows a plot of the likelihood scores for 50,000 people. See [14] for analysis results.

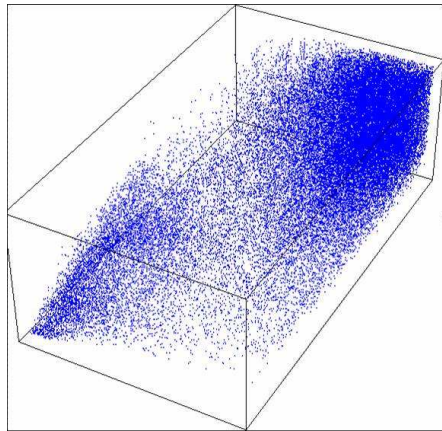


FIGURE 3. Likelihood scores for neighborhoods.

The goal in this particular project is to identify people with small likelihood scores, based on multidimensional order statistics, as potential candidates for intervention. Details will appear in [14].

4. DISCUSSION

In this paper we have introduced a quantitative method for the analysis of dynamic network observations. From a trajectory of observations of interactions we derive a collection of measures that capture the connectivity properties over time of the neighborhood of each node in the network. That is, we consider all nodes that interact with a given node over time, as well as the interactions of those nodes with each other. We point out that while our basic approach is similar to the study of "communities of interest" as proposed in, e.g., [6, 7], the notion of "community" there is limited to a given node and its outgoing and incoming edges, and does not consider the topology of the "community" as a whole.

Our method provides characteristic measures for each network node that can be used for comparison of nodes, clustering approaches to find nodes with "unusual" neighborhood dynamics, as well as the study of distributions of neighborhood characteristics. Our software package, available from the second author, makes feasible the analysis of networks with millions of nodes and thousands of observations.

REFERENCES

- [1] R. Atkin, An Algebra for Patterns on a Complex, I, *Internat. J. Man-Machine Stud.* **6**, 285-307, 1974.
- [2] R. Atkin, An Algebra for Patterns on a Complex, II, *Internat. J. Man-Machine Stud.* **8**, 483-448, 1976.
- [3] H. Barcelo, R. Laubenbacher, Perspectives on A-homotopy theory and its applications, to appear, special Issue of *Discr. Math.* for FPSAC 2002.
- [4] H. Barcelo, X. Kramer, R. Laubenbacher, and C. Weaver, Foundations of a Connectivity Theory for Simplicial Complexes, *Adv. Appl. Math.* **26**, 97-128, 2001.
- [5] E. Babson, H. Barcelo, M. De Longueville, and R. Laubenbacher, Homotopy theory of graphs, *J. Alg. Comb.*, to appear.
- [6] C. Cortes, D. Pregibon, and C. Volinsky, Communities of interest, preprint, 2004.
- [7] C. Cortes, D. Pregibon, and C. Volinsky, Computational methods for dynamic graphs, preprint, 2004.
- [8] S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, Z. Toroczkai, and N. Wang, Modelling disease outbreaks in realistic urban social networks, *Nature* **429** (2004) 180-184.
- [9] R.A. Fabes, S.A. Shepard, I.K. Guthrie and C.L. Martin, *Roles of temperamental arousal and gender-segregated play in young children's social adjustment*, *Developmental Psychology*, **33**, (1997), 693-702.
- [10] W.A. Griffin, L.D. Hanish, C.L. Martin and R.A. Fabes, *Modeling playgroups in children: Determining validity and veridicality*, In D. L. Sallach, C. M. Macal, M. J. North (Eds.), *Agent 2003: Challenges in social simulation* (pp. 93-111). Chicago: University of Chicago and Argonne National Laboratory, 2004.
- [11] Hanish Laura D., Martin Carol Lynn, Fabes Richard A., Barcelo Hélène, Griffin William, Schmidt Shana, Dodd Mike, *Methodological Advances in Studying Peer Relationships: the Q-Connectivity Approach*, in preparation.
- [12] J. Johnson, *The Mathematics of Complex Systems*, in *The Mathematical Revolution Inspired by Computing*, Oxford University Press, Oxford, 1991.
- [13] C.L. Martin, and R.A. Fabes, *The stability and consequences of young children's same-sex peer interactions*, *Developmental Psychology*, **37**, (2001), 431-446.
- [14] J. McGee, H. Barcelo, and R. Laubenbacher, Q-analysis methods for the study of dynamic contract graphs of social networks, in preparation.
- [15] Newman, M. E. J. (1-MI-P) *The structure and function of complex networks*, *SIAM Rev.* **45** (2003), no. 2, 167-256.
- [16] Wasserman S. and Faust K. *Social network Analysis: Methods and Applications*, Cambridge U. Press, 1994.

DEPARTMENT OF MATHEMATICS AND STATISTICS, ARIZONA STATE UNIVERSITY, TEMPE, ARIZONA 85287-1804
E-mail address: `barcelo@asu.edu`

VIRGINIA BIOINFORMATICS INSTITUTE, VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY, BLACKSBURG, VA 24061
E-mail address: `reinhard@vbi.vt.edu`