Title: Querying Volatile and Dynamic Networks
Name: Sarvenaz Choobdar   Pedro Ribeiro   Fernando Silva
Affil./Addr.:   CRACS and INESC-TEC
                Faculdade de Ciencias, Universidade do Porto, Portugal
                Email:{sarvenaz,pribeiro,fds}@dcc.fc.up.pt

Querying Volatile and Dynamic Networks

Synonyms

Network Dynamics, Structural Roles, Clustering, Event Detection

Glossary

Dynamic (or volatile) network: a network whose structure changes over time, some nodes and edges may appear and disappear.

Local property: a metric that describes a topological aspect on the local neighborhood of a node.

Structural role: the structural position of a node in the network, characterized by its local properties.

Transition pattern: a typical role change, characterized by its time interval and by the origin and destination role.

Definition

Many networks are intrinsically dynamic and change over time. These networks can be very volatile, with a significant number of edges and nodes appearing and disappearing. The majority of the existing network mining methodologies are however geared towards a more static scenario, with a single graph describing the topology of the system being analyzed. There is still a need for measurements and tools that allow the temporal dimension on network analysis to be fully and naturally incorporated.

In this paper we deal precisely with volatile and dynamic networks, and how to extract information from them. We take a general approach in which we track the structural role of nodes over time. By using a set of local properties, such as degree and clustering coefficients, we group similar nodes and we monitor the behavior of these groups over time, looking for significant changes. We are able to describe the evolution of the network in the form of transition patterns that explain how position of nodes change and how groups of nodes grow and shrink.

Introduction

Networks are a powerful and flexible model to represent a multitude of systems. Research on network analysis in general, and social networks in particular, has been very prolific. There is a large number of measurements available for a practitioner wanting to understand a particular network. For a good overview see Costa et al [5].

The methodologies used have a broad scope and span many different conceptual metrics. Methods such as community detection [7] tackle potentially large groups of nodes. Others concentrate on smaller groups of nodes, such as network motif discovery [20] or frequent subgraph mining [12]. Finally, there are methodologies that try to analyze at the individual level, for instance by classifying nodes [19].

Here we focus on this last angle of approach and analyze the network from the point of view of individual nodes by measuring local properties of nodes such as clustering coefficient [21] or the degree distribution [6]. Our main focus is on the temporal dimension of networks to discover how structural characteristics of nodes change over time. We use a set of local properties that fully characterize the topology of single nodes to define structural positions of nodes, and we track how these positions evolve in networks lifespan. In order to better understand the global network dynamics, we group the nodes based on structural properties and we observe the temporal behavior of nodes. In this paper we are trying to answer questions such as: What structural positions or roles do exist in a network? How are they appearing and disappearing? Are specific roles growing or shrinking? Which periods of time show significant changes and which groups of nodes exemplify such changes? Moreover, we are also interested in understanding the transitions between roles of nodes. Which role is losing nodes to another? What is the typical path taken by the structural role of nodes? Can we pinpoint in time and identify major role transitions?

Key Points

Our proposed methodology is geared towards a dynamic network that keeps a relatively static set of nodes (so that we can track them over time), but has a volatile set of edges, that keep changing and evolving. We are able to characterize single nodes using local properties, to group them by structural function using clustering over the properties, and to discover significant events, positioning them in time by discovering the relevant time intervals, and identifying the relevant transition patterns by creating sets of rules that describe the change of nodes from one role to another.

Historical Background

Only recently, more attention is being given to the dynamic aspects of networks. From a global perspective, research has been devoted to the study of general growth patterns of networks [16; 15; 25], and, on a less global scale, to aspects such as the formation

and evolution of communities [9; 18; 26], or subgraph dynamics [1; 14]. On a more local level, link prediction has also been used to guess appearence of edges over time [17].

Modeling the temporal behavior of nodes is a relatively emergent point of view. In this line of research the main focus is on the single nodes and their properties in the network. In our previous works [2; 3] a two-phase general methodology was designed to characterize time evolving networks. In the first step of this methodology nodes are grouped by k-means clustering and classified based on their role in the network. In the second step a method is proposed to study the evolution of the network by a supervised approach. In this method a set of events happening in the network is defined for the roles in the network. We then find the predefined events happening in the network and the rules that describe them by using association rule mining. On a related approach, Rossi et al [22] used the methodology proposed by [11] for static role extraction. They measure a set of features for nodes at each time snapshot then by stacking all the node-by-feature matrices, they derive the matrix of feature-roles by factorizing the stacked node-by-feature matrix and iteratively generate the matrix of node-role for each time.

Proposed Methodology

In this section we formulate the problem of dynamic network study from a single node point of view. Our goal is to model the temporal characteristics of nodes regarding the different structural positions they hold, which we call *roles* in this paper. For a given sequence of network snapshots, we examine the dynamics by tracking roles of nodes over time. For a given node, the role is defined regarding properties in its close neighborhood. The proposed method for this problem consists of three steps: (1) measuring local properties of nodes at each time step; (2) extracting structural roles of nodes; (3) finding patterns of role transition.

Measuring Local Properties

The role of a node is automatically determined based on its properties in the network. Nodes are assigned to the same role if they are similar. We assess the similarity of two nodes by using their properties rather than number of edges between them. Two nodes are close if they have a similar *feature vector.* Initially, we need to select a set of local measurements that best characterize nodes in the network structure. We chose to employ the same set of metrics used by [4] as the feature vector for finding groups of nodes. These features measure the connectivity of a node in the neighborhood structure. In particular, we use the following five properties:

- the normalized average degree (r);
- the coefficient variation of the degrees of the immediate neighbors of a node (cv);
- the clustering coefficient (cc);

- the locality index (loc), which is an extension of the matching index and takes into account all the immediate neighbors of each node, instead of individual edges, and
- the normalized node degree (K).

Extracting Structural Roles

To extract the set of roles describing the existing positions in the networks, we use the well known k-means clustering algorithm [10] to find groups of nodes with similar feature vectors. Each cluster contains nodes with a similar position in the network regarding their feature vectors, hence the same role or label can be assigned to them.

Given a sequence of graphs $\{G_1, G_2, ..., G_t\}$, we build the feature matrix M_t by calculating the local properties of nodes at time t. We discover groups of nodes by applying k-means algorithm on the stacked matrix of all node-feature matrices up to time t. In the stacked matrix, one node may appear multiple times with different feature vectors, if it is active (i.e., if it exists) at multiple time steps. We derive a membership matrix (node by time) for all nodes of the network regarding the clustering result that shows the role of nodes at each time step.

The number of potential groups of nodes in the network is equal to the number of clusters in the dataset. Determining the actual number of groups in a dataset is a fundamental and largely unsolved problem in cluster analysis. We employ the method by [24], since it does not require parametric assumptions, is independent of the method of clustering, and was shown to achieve excellent results. This method uses a theoretic information approach that considers the transformed distortion curve $d : K^{-p/2}$, where p is the number of dimensions in the dataset.

Role transition modeling

We explore the dynamic of a network at node level by finding role transition patterns. The temporal behavior is modeled to see if nodes hold similar roles over their life time or consistently oscillate over time. For example, in a social network, a center of a star was a bridge node before or is it likely to turn to an isolated node in its life span? We generate a set of transition rules to describe these events in a network.

Five basic types of temporal patterns are defined for a role regarding the number of nodes having that role: (1) growth; (2) shrink; (3) emergence; (4) dissolution; (5) constant. We learn these patterns by discovering transition intervals where the number of nodes of a certain role is shrinking, growing or showing other defined temporal patterns. After this, a set of rules describing the pattern is generated.

Transition intervals

A transition interval is a time interval where a considerable number of nodes leave or join a role. For a given role C_i, its size over time constitutes a time series denoted as $F_i(t), t \in [1, T]$. A transition interval is the subsequence of F_i which holds a constant

increasing or decreasing trend. Hence, we extract transition intervals of F_i by segmentation of the time series. Starting from $t = 1$, $F_i(t)$ is approximated by linear regression to find the transition intervals. If the error of the fitted line for a subsequence $F_i[a : b]$ exceeds the threshold, the interval $[a, b]$ breaks to point j where it gives the best approximation for $F_i[a, j - 1], j < b$. The error is measured in terms of the sum square of residuals. The threshold is controlled by the maximum number of arbitrary transition intervals. The maximum error is increased while the number of intervals is below the defined maximum number of intervals. For this method we can either define the maximum error for the linear regression, or the maximum number of desired intervals. The slope of the fitted line for each segment shows whether the interval is increasing or decreasing, which respectively determines the growth (emergence) or shrink (dissolution) events.

Transition rules

We extract a set of rules to describe the specific pattern at transition intervals. A transition rule is in the form of $C_i \rightarrow C_j$, which shows that nodes from group i moved to j. The order of transitions is important, since it shows the trend of changes in the network properties. For example, for a sequence of $\{1, 1, 1, 2, 2, 3\}$, the transition rules are $1 \rightarrow 2$ and $2 \rightarrow 3$. We extract these one-step transition rules for a time interval by building a transition matrix in each time interval. The support count of a rule $C_i \rightarrow C_j$ is defined as the number of nodes that go from cluster C_i to C_j in that interval.

Exploratory analysis

We now apply our proposed methodology to real data in order to show its applicability. For our experiments, we used three different real complex networks: world countries global trade [8] (two countries A and B are connected at time t if the trade between them is more than 10% of either country's total trade in that year), connections between USA airports [23] and a co-authorship network obtained from DBLP data [1]. We use the undirected forms of these networks. Table 1 gives some more details on the topology of three used networks.

Dataset	Time snapshots	$|V|$	$|E|$	Node growth rate	Edge growth rate
Global Trade	53	186	8839	2.47	7.93
USA Airports	244	1919	14391	1.64	1.21
DBLP	11	31592	49599	3.4	4.57

Table 1: Datasets statistics: number of time-snapshots, number of nodes and edges at the final snapshot and node and edge growth rate (ratio between the final and initial time-snapshots).

This methodology gives a profile for each network, including information about the existing roles and the temporal patterns, describing the existing dynamics. Figure 1 shows the profile of the roles found in each network. The profile depicts the values of the feature vector of each role in the network. This figure clearly depicts that each role in the network has a different feature vector and that nodes are distinguished according to their structural position.
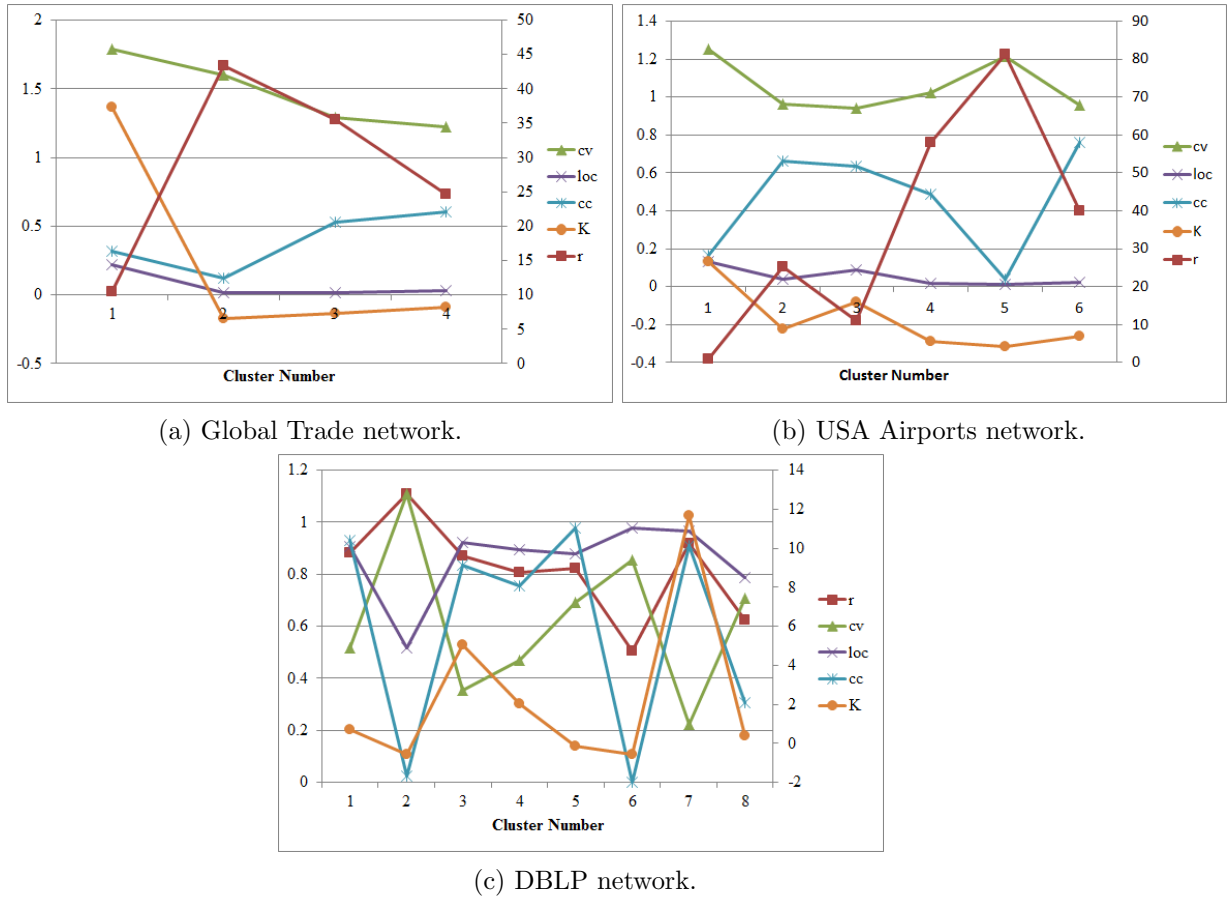

(a) Global Trade network.

(b) USA Airports network.

(c) DBLP network.

Fig. 1: Illustration of the feature vector of the roles in the selected networks. The normalized average degree (r) is plotted on the second axes on the right side of the plot.

For example, in the global trade network of countries, our method found four distinguishable roles for nodes. Nodes with role number 1 are countries with very high degree and many low degree nodes connected to them. Neighbors of these nodes have low degree since the normalized average degree of the immediate neighbors of a node for this group is very low. This means that nodes in role 1 behave as hubs in the network, i.e., as hub countries in global trade, with commercial transactions with many other countries that have a high variation of degree in neighborhood (cv). According to the value of (loc) and (cc), respectively, the locality index and the clustering coefficient, nodes are highly connected in their neighborhood. United States of America, Canada and France are members of this group.

Dataset	Event	Time interval	Transition rules	Support	Z score
Global Trade	Emerge (2)	[26,32]	$3 \to 2$	80	6.3
	Shrink (4)	[23,40]	$4 \to 3$	100	14.70
	Emerge (3)	[12,22]	$4 \to 3$	126	8.3
	Shrink (1)	[9,29]	$1 \to 4$	25	12.1
USA airports	Growth (3)	[142,175]	$1 \to 3$	1062	3.75
	Growth (6)	[1,79]	$2 \to 6$	437	7.27
	Shrink (2)	[57,132]	$2 \to 3$	551	4.79
DBLP	Growth (4)	[1,11]	$0 \to 4$	6710	1.2
	Growth (7)	[6,11]	$0 \to 7$	196	0.2
	Shrink (3)	[1,7]	$3 \to 0$	726	1.3

Table 2: Description of some extracted events in the networks. Numbers in the parenthesis denotes the cluster number, holding the events

Table 2 shows some of the extracted patterns. Besides the frequency count of every pattern in each dataset, and in order to assert their significance, we compared the result with randomized sequences. We built these random sequences by shuffling the order of role membership of each node in the network. In this way, the random dataset has the same number of nodes and role types. Since we have a dataset of role sequences for each network, we built 10 datasets of random sequences with the size as the number of nodes. We calculated the average and standard deviation of frequencies for each rule of events in the 10 datasets. Finally, we computed the z-score for the significance of each rule as compared to the randomized form. The table shows the rules with the highest z-score.

Figure 2 graphically illustrates some of the role transition patterns happening. Each graph includes all roles transitions in a specified time interval. The label of nodes specifies the role and the color of nodes shows the type of temporal pattern.

For example, the first event (a) in the global trade network describes the emergence of cluster 3 in the [12,22] time interval. This graph says that nodes that constitute cluster 3 either come from cluster 4 or are the new nodes (cluster 0), just joining the network. By observing Table 2, we can see that the main reason for this event is the transition from cluster 4. The same type of interpretation could be applied for the two other networks.

Regarding the scalability of our approach, out methodology gives results in a reasonable time for small scale networks like Global Trade or a relatively large network like DBLP. The size of the network essentially affects the first step of the proposed methodology where the local properties of nodes are calculated. In this paper, this set of properties can be computed efficiently since they are based on one 1st degree ego network of nodes. The other steps of the method utilize standard machine learning techniques. The clustering step uses K-means, which is an NP-hard problem, but there are efficient approximate solutions [13]. The role transition modeling step is more af-
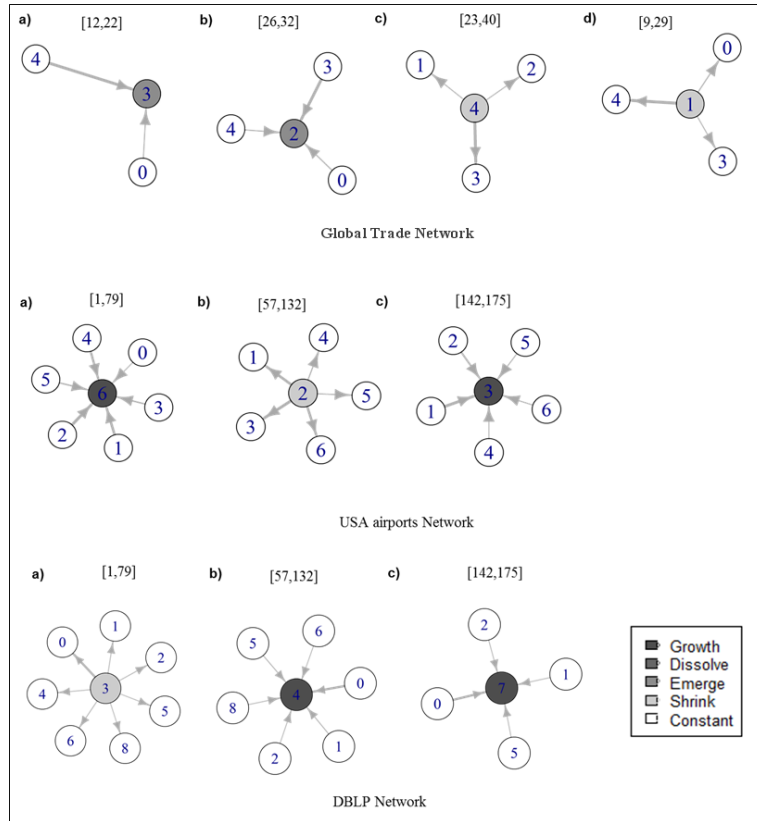
Fig. 2: Some of the events happening in the networks.

fected by the number of time snapshots rather than the size of network since transition rules are derived by matrix operations in every time step.

Key Applications

This method provides a general picture of a network and its dynamics over time, and as such has broad applicability. The results of our methodology are helpful for comparing networks regarding their role distribution. For example, we can examine if there is the same role distribution in social and biological networks. Finding outliers in a network is another possible application. Nodes that exhibit a temporal behavior different from the other nodes in their discovered role set can be considered as outliers.

Future Directions

In the near future we indent to study the usage of other types of measurements. For instance, can subgraph metrics, such as participation of nodes in network motifs, be used for role discovery? We also intend to study the applicability of our methodology to very large scale networks, by scaling up our used algorithms. In addition, we plan to apply the proposed method for network studies such as network comparison, node classification and anomaly detection.

Cross-References

- Analysis and Visualization of Dynamic Networks
- Community Identification in Dynamic and Complex Networks
- Future Directions in Predicting the Dynamic Evolution of a Network
- Social Networks: Dynamics and Evolution Models for
- Temporal Analysis on Static and Dynamic Social Networks Topologies
- Role Discovery
- Role Identification of Social Networkers
- Path-Based and Whole Network Measures

Acknowledgements

References

1. Berlingerio M, Bonchi F, Bringmann B, Gionis A (2009) Mining graph evolution rules. In: Machine Learning and Knowledge Discovery in Databases, Springer, pp 115–130
2. Choobdar S, Silva F, Ribeiro P (2011) Network node label acquisition and tracking. In: Progress in Artificial Intelligence, 15th Portuguese Conference on Artificial Intelligence - EPIA'11, Lisbon, Springer, LNAI 7026, pp 418–430
3. Choobdar S, Silva F, Ribeiro P (2012) Event detection in evolving networks. In: International Conference onComputational Aspects of Social Networks (CASoN), 2012
4. Costa L, Rodrigues F, Hilgetag C, Kaiser M (2009) Beyond the average: detecting global singular nodes from local features in complex networks. EPL (Europhysics Letters) 87:18,008
5. Costa LF, Rodrigues FA, Travieso G, Boas PRV (2007) Characterization of complex networks: A survey of measurements. Advances In Physics 56:167
6. Dorogovtsev SN, Mendes JFF (2005) The shortest path to complex networks. Complex Systems and Inter-disciplinary Science 1
7. Fortunato S (2010) Community detection in graphs. Physics Reports 486(3-5):75 – 174
8. Gleditsch K (2002) Expanded trade and GDP data. Journal of Conflict Resolution 46(5):712
9. Greene D, Doyle D, Cunningham P (2010) Tracking the evolution of communities in dynamic social networks. In: Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on, IEEE, pp 176–183
10. Hartigan J, Wong M (1979) A k-means clustering algorithm. Journal of the Royal Statistical Society C 28(1):100–108
11. Henderson K, Gallagher B, Eliassi-Rad T, Tong H, Basu S, Akoglu L, Koutra D, Faloutsos C, Li L, Matsubara Y, et al (2012) Rolx: Structural role extraction & mining in large graphs. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 1231–1239
12. Huan J, Wang W, Prins J (2003) Efficient mining of frequent subgraphs in the presence of isomorphism. In: Proceedings of the Third IEEE International Conference on Data Mining, ICDM'03, pp 549–552

13. Jain AK (2010) Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31(8):651–666

14. Jin R, McCallen S, Almaas E (2007) Trend motif: A graph mining approach for analysis of dynamic complex networks. In: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, IEEE, pp 541–546

15. Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining, pp 177–187

16. Leskovec J, Backstrom L, Kumar R, Tomkins A (2008) Microscopic evolution of social networks. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 462–470

17. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. Journal of the American society for information science and technology 58(7):1019–1031

18. Lin Y, Chi Y, Zhu S, Sundaram H, Tseng B (2008) Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In: Proceedings of the 17th international conference on World Wide Web, ACM, pp 685–694

19. Macskassy S, Provost F (2007) Classification in networked data: A toolkit and a univariate case study. The Journal of Machine Learning Research 8:935–983

20. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network Motifs: Simple Building Blocks of Complex Networks. Science 298(5594):824–827

21. Newman ME (2001) Scientific collaboration networks. I. Network construction and fundamental results. Phys Rev E Stat Nonlin Soft Matter Phys 64(1 Pt 2)

22. Rossi R, Gallagher B, Neville J, Henderson K (2012) Role-dynamics: fast mining of large dynamic networks. In: Proceedings of the 21st international conference companion on World Wide Web, pp 997–1006

23. RouteViews (1997) University of Oregon Route Views Project. Online data and reports. `http://www.routeviews.org` (accessed February 2013)

24. Sugar C, James G (2003) Finding the number of clusters in a dataset. Journal of the American Statistical Association 98(463):750–763

25. Sun J, Tao D, Faloutsos C (2006) Beyond streams and graphs: dynamic tensor analysis. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 374–383

26. Tang L, Liu H, Zhang J, Nazeri Z (2008) Community evolution in dynamic multi-mode networks. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 677–685