Using implications from FCA to represent a two mode network data

Sebastião M. Neto, Mark A. J. Song,
Luis E. Zárate
Centro Universitário UNA
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte – MG – Brasil
mark@prof.una.br,
sebastiaomendesneto@gmail.com,
zarate@pucminas.br

Sergio M. Dias
Serviço Federal de Processamento de Dados
SERPRO
Belo Horizonte – MG – Brasil
sergio.dias@serpro.gov.br

Abstract* – *In a world of ever-growing connectivity, full of connections between people and objects, new multidisciplinary complex network analysis needs to arise. This work presents a solution to analyze an Internet Service Provider database using a formal concept analysis element named implications and complex network techniques. Our goal is to analyze access to the 25 most visited websites to find access patterns. We selected 9 time intervals in one week. Data were converted to a clarified formal context and the FindImplications algorithm was used to extract implications sets. These sets were cross-checked to look for patterns. The implications were used to explore the complex network substructures. As a result, we found access patterns that guarantee that whenever premise websites are accessed, so are conclusion websites. This result can aid in creating security policies and network configurations to help predict future accesses. Without this technique relationships between events nodes (websites) of a two mode network could not be identified.

I. Introduction

In the last few years, attention has turned towards the increasing complexity of the connected world, as noted by Easley and Kleinberg [1]. This connectivity is propelled by a variety of factors, such as the Internet itself, telephone networks and the speed with which information travels around the globe. These factors enable the genesis of social networks formed by relationships between people. Motivated by the interconnected world, research has surfaced and disciplines interlink to contribute with techniques and new perspectives for the analysis of these complex networks.

Currently, social network analysis is focused on the discovery of social relationship patterns. These relationships can occur between subjects, events, or subjects and events. According to Getoor and Diehi [2], in some cases some relationships are not observed. Therefore, it is of general interest to unveil hidden substructures as possible and potential communities.

However, the identification of substructures demands work towards perfecting methods to clarify network visualization and extract representations and important knowledge from them [3,4]. We propose a implications-based computational models that allows for the extraction of new knowledge and better visualizations.

The knowledge extraction, in turn, are done via Formal Concept Analysis (FCA) [5], which is a mathematical research field introduced by Rudolf Wille and has found use in different fields. Due to its potential in knowledge representation, FCA can also benefit complex network analysis, as discussed in related work section.

In this work, we used implications seeking to increase our knowledge on complex networks and innovates by using implications to find patterns, build graphs and conduct analyses in a two mode network data.

II. Complex Networks and Limitations

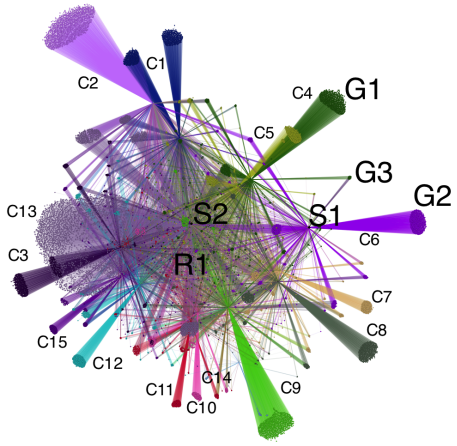
Graph theory [6] is a framework which enables the representation of complex networks in a mathematically accurate way. Formally, a network can be represented by a graph $G = (N, L)$ directed or not, where $N \neq \emptyset$ and L are sets of pairs, sorted or not, of elements from N. The elements in $N \equiv \{n_1, n_2, \ldots, n_M\}$ are vertices of graph G. The elements in $L \equiv \{l_1, l_2, \ldots, l_K\}$ are the edges.

Networks that contain vertices of a single type are called one mode network data. When networks contain vertices of two or more types, such as vertices that represent users and websites, they are called two mode network data [7].

Fig. 1 shows an example of a network that represents user connections to websites. The resulting graph is a bipar-

(DOI reference number: 10.18293/SEKE2015-085)

Figure 1: Network generated from the accesses of 20.115 users in a day to 15 preselected websites.



tite directional graph, with vertices representing users connected to vertices representing visited websites. It is a two mode network data. A cluster finding algorithm, *Clustering Chinese Whispers* [8], found 15 groupings (C1-C15). These users have a greater probability of communicating and constituting a social network [9][10]. In the groupings G1 and G2, each user accessed a single website. Therefore, they are so next to each other [11]. The grouping G2 accessed the website S1. Users of grouping G3 accessed two websites and, therefore, is positioned between G1 and G2. Website S2 (Google), located in the center of the graph, have a vast number of user connections and its centrality corroborates its importance in the graph. The Region R1, made up of users who accessed many websites, ended up not grouped and becoming dispersed.

Relationships between websites are not easily observed. This begs the question: how to determine which access to a set of websites implies in an access to another website? User-website relationships are perceptible, they are represented by graph edges. However, inter-website relationships cannot be inferred. Freeman and White [12] exposed this limitation in the representation of networks containing nodes of two types (two mode network data) and suggested the concept lattices as the best option. In our work, we showed that implications along with complex network techniques can also help understand this kind of network.

III. Formal Concept Analysis and APRIORI

FCA [5] offers the formalization of a concept, which is made by an intention and an extension. The extension corresponds to all objects belonging to the concept, while the intention represents all attributes shared by the aforementioned objects. FCA allows us to identify object groups with a specific meaning that share common attributes. According to Ganter and Wille [5], FCA revolves around four fundamental elements: formal contexts, formal concepts, concept lattices and implications.

A formal context is usually represented by a table, in which rows represent objects and columns represent attributes. When an object possesses an attribute we have an incidence represented by an "X" [5].

Formally, a formal context has the notation $K := (G, M, I)$, where G is a set of objects (rows), M is a set of attributes (columns) and I are incidences, defined as $I \subseteq G \times M$. If an object $g \in G$ and an attribute $m \in M$ are in relationship I, their representation is $(g, m) \in I$ ou gIm, which reads as "object g has attribute m".

Given a subset of objects $A \subseteq G$ of a formal context $K := (G, M, I)$, there is an attribute subset of M common to every object of A, even if empty. Likewise, given a set $B \subseteq M$, there is an object subset that shares the attributes of B, even if empty. These relationships are defined by derivation operations [5]: $A' := \{m \in M | gIm \forall g \in A\}$ and $B' := \{g \in G | gIm \forall m \in B\}$

From the formal contexts we obtain formal concepts, defined as pairs (A, B), where $A \subseteq G$ is called extension and $B \subseteq M$ and is called intention, and they must follow the conditions $A = B'$ and $B = A'$ [5].

With all formal concepts sorted hierarchically by order of inclusion $\subseteq$ we can build the concept lattice.

Implications exist between two attribute subsets of a formal context. Formally, an implication can be expressed as follows. Considering a context $K := (G, M, I)$ satisfying implications $Q \rightarrow R, Q, R \subseteq M$, if for every $g \in G, gIq$ for every $q \in Q$ implies gIr for every $r \in R$. These implications are normally used in data mining to find dependencies [4].

It is important to note that association rules obtained by the famous APRIORI [13] algorithm are usually expressed with a confidence interval of 0% to 100%. Implications based on existing formal concepts always yield a confidence rate of 100%. Therefore, depending on what the attribute represents, like websites, we can establish an access pattern to related websites.

IV. Related Work

Poelmans et al [14] presents a "semiautomatic" process to expose a network of criminal organizations and their members. They built a lattice of suspected drug dealers and then a lattice containing suspect profiles, that allowed the identification of criminal networks.

Freeman [15] found several important complex network analysis elements through lattice observations, like cliques and bridging cliques, showing that these elements can be observed by FCA and used to facilitate the analysis of complex networks.

Cuvelier and Aufaure [16], analyzed tweets about a specific subject. By means of FCA, the authors were able to establish relationships between messages and, after representing them in the lattice, use data filtering criteria to assemble a topographical network graph to help clarify the information obtained.

Aufaure and Le Grand [17], who describe lattice expressiveness, especially when associated with ontologies, as a benefit of this FCA use. The work is a compilation of case studies. Among the conclusions, they have observed that lattices allow researchers to find deterministically-overlapping groupings, which can be labeled using extensions and intentions.

Our work differs by searching implications for a way to provide more knowledge about the influence of event type vertices in two mode network data networks when looking for substructures.

V. METHODOLOGY

Freeman [12] mentioned the problem of finding the relationships between event type vertices in two mode network data. Here, this problem is attacked by transforming network data into a formal context, extracting implications and finally make a network composed by sets of websites (premises and conclusions) connected by edges (implications). This way, we turn a two mode network data into a one mode network data. With that, we expose existing relationships between event type vertices, which was not feasible in the original network. This would render every complex network technique applicable to the analyzed networks.

The FindImplications [18] algorithm, gather from [19], was used to extract implications. It takes a formal context as input and looks for an implication coverage. It is known for its completeness, being capable to extract all implications.

The object clarification process consists in eliminating duplicated objects. Objects with the same attribute sets can be discarded when obtaining implications, since they do not influence the result [5]. So, we have only considered clarified contexts.

VI. EXPERIMENT AND RESULT ANALYSIS

The database used was provided by a Brazilian cable Internet Service Provider, with anonymous access data. Records refer to accesses made in march 2009, adding up to a total of 6,319,333 accesses.

Targeting the 25 most accesses websites, we found 165,659 (24,9% of the overall access total) accesses made by 29,319 distinct users in the first week of the month. The aiming was to find access patterns among week days. The data were converted to a formal context with websites as attributes and users as objects. Websites from the same domain, for example "www.globo.com", "ads.globo.com"and "bbb.globo.com" were grouped together for simplification sake, and categorized only as "globo". The number of attributes was reduced to 15.

Nine formal contexts were generated. Table I shows contexts, the time of day in which accesses were logged, the number of users in the context (objects) and the number of profiles (user sets with identical accesses) in the clarified context.

TABLE I. CONTEXTS AND THEIR CHARACTERISTICS

Context	Access Period	Users	Profiles
K1	Monday from 8 AM-6 PM	11.387	715
K2	Tuesday from 8 AM-6 PM	11.095	710
K3	Wednesday from 8 AM-6 PM	10.829	672
K4	Thursday from 8 AM-6 PM	10.860	700
K5	Friday from 8 AM-6 PM	10.633	683
K6	Friday from 10 PM-2 AM	5.445	315
K7	Saturday from 10 PM-2 AM	5.172	294
K8	Sunday from 10 PM-2 AM	5.482	309
K9	Wednesday from 10 PM-2 AM	5.487	296

The amount of implications extracted and the number of intersections between contexts are shown in Table II.

TABLE II. IMPLICATIONS AND INTERSECTIONS

Cxt	N.R	K2	K3	K4	K5	K6	K7	K8	K9
K1	1.246	50	35	20	51	0	2	0	0
K2	876		29	13	19	2	2	0	0
K3	955			37	60	0	5	0	1
K4	835				46	0	2	0	0
K5	919					1	0	0	1
K6	205						0	0	1
K7	123							0	3
K8	155								0
K9	149								

In addition to the aforementioned intersections, other intersections were found, like between K1, K2, K3 and K4, and the following implication rule, common to all contexts, was also found:

$$\{orkutgstatic, ad.doubleclick.net, yahoo\} \rightarrow \{google\} \quad (1)$$

In rule (1), with 100% certainty, we identified an access pattern which repeated itself in 4 days of the week, from Monday to Thursday. These websites are theoretically harmless. However, in case of a rule which exposes improper behavior, with a premise website set that leads to dangerous conclusion websites, it can be used as a malicious behavior pattern and have alerts and special controls associated to it.

The support for a rule is the percentage of objects that follow the rule relative to the total amount of objects. For rule (1), the approximate support percentage was approximately 0.11% for the non-clarified context K1. Thus, this rule identifies 13 users. For non-clarified context K2, with 11,095 users, the rule has an approximate support rate of 0.08%, identifying 9 users. Context K3, which has 10,829 users, yielded a support rate of 0.06% for the rule, identifying 7 users. For context K4, the support rate was also approximately 0.06%, and the number of identified users was also 7 since K4 has 10,860 users. With this access pattern, we reached a total of 32 distinct identified users, of which four followed the pattern in at least two days.

If we take rule (1) as suspicious user behavior, we could start a detailed control process, tailored to recurring users. As soon as one of them or, if desired, one of the 28 others accesses a premise website, a security policy can be initiated.

The network formed by the implications yields a bipartite, directed graph, with edges going from the premise to the conclusion.

Figure 2: Approximate depiction of the network formed by implications

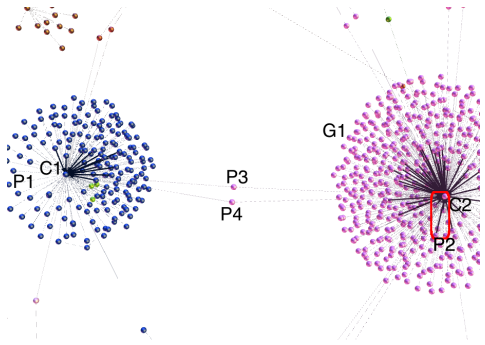


Fig. 2 shows an approximate depiction of a network formed by implications from contexts K3 and K5. Groupings, like G1, indicate premises (websites) with equal conclusions. The algorithm *Clustering Chinese Whispers* identified 449 groupings, implications with the same conclusion. Vertex P1 represents a premise linked to a conclusion C1 (google, yahoo). Vertex P2 is linked to conclusion C2 (google), forming an implication that was recurrent in both days (darker edge in the highlighted area). By making use of strong ties principles [1], it is possible to determine that premises linked to the same conclusion represent users who have similar behavior. If the access to conclusion sites raises concern, like increased bandwidth consumption, when one of the premises occurs, automatic configurations to avoid network bottlenecks can be triggered. Premises P3 and P4 are linked to more than one conclusion, indicating ambiguous behavior for the premises. This only took place because the graph is formed with implications from two days. None of these observations, made from the network composed by premise and conclusion website sets, would have been possible if the original data was visualized and analyzed.

VII. Conclusions

The main goal of this work was to associate FCA and complex networks to allow us to uncover access patterns (recurrent and ambiguous), in addition to enabling a better representations of relationships between event type elements in two mode network data complex networks.

Implication obtained from user connections made the discovery of access patterns viable. By obtaining rule intersections, patterns and recurring users were identified. The network formed by the rules found ambiguous premises and premise groups with the same conclusion. Examples of use for this type of information were presented, like automatic safeguarding measures to improve service provided and security controls.

We intent to add another dimension, time, to collected data for access prediction analyses based on a profile containing access sequences. For highly dimensional formal contexts, execution times of the algorithms become prohibitive, so we intent to find optimizations for these algorithms using new formal context representations (such as Binary Decision Diagrams and parallelism).

Find more information about this work at http://goo.gl/n2G6lI.

VIII. Acknowledgments

We thank SERPRO and also FAPEMIG, CNPq and CAPES for their financial support.

References

[1] D. Easley and J. Kleinberg, “Networks, crowds, and markets: Reasoning about a highly connected world”, Wiley Online Library, 2012.

[2] L. Getoor and C. Diehl. “Link mining: a survey’, ACM SIGKDD Explorations Newsletter, Vol. 7, Issue 2, pp. 3-12, 2005.

[3] L. C. Freeman, “Visualizing social networks”, Journal of social structure, vol. 1 2000.

[4] L. C. Freeman, “Graphical techniques for exploring social network data”, Models and Methods in Social Network Analysis, 2005.

[5] B. Ganter, G. Stumme, and R. Wille, “Formal concept analysis: foundations and applications”, Dresden, Alemanha, Springer, v.3626. 2005.

[6] B. Bollobas, “Random graphs”, Academic Press, London, 1985.

[7] S. Wasserman and K. Faust, “Social network analysis: methods and applications”, New York. Academic Press. 1993

[8] C. Biemann, “Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems”, In Proceedings of the first workshop on graph based methods for natural language processing, p. 73-80, Association for Computational Linguistics, 2006.

[9] L. B. Ronald, “The duality of persons and groups”, Social Forces, vol. 53, pp. 181-190, 1974.

[10] S. L. Feld. “The focused organization of social ties”. American Journal of Sociology, 86(5), pp. 1015-1035, 1981.

[11] M. Jacomy, S. Heymann, T. Venturini, and M. Bastian, “Forceatlas2, a graph layout algorithm for handy network visualization”, Paris http://www. medialab. sciences-po. fr/fr/publications-fr , 2009.

[12] L. C. Freeman and D. R. White, “Using galois lattices to represent network data”. Sociological methodology, v. 23, pp. 127-146. 1993.

[13] R. Agrawal, S. Ramakrishnan, “Fast algorithms for mining association rules”, In Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, pp. 487-499, 1994.

[14] J. Poelmans, P. Elzinga, S. Viaene, G. Dedene, and S. Kuznetsov, “A concept discovery approach for fighting human trafficking and forced prostitution”, 19th International conference on conceptual structures, lecture notes in computer science, vol. 6828, pp. 201-214, Derby, England: Springer, 2011.

[15] L. C. Freeman, “Cliques, galois lattices, and the structure of human social groups”, Elsevier, Social Networks, 18, pp. 173-187, 1996.

[16] E. Cuvelier and M. Aufaure, “A buzz and e-reputation monitoring tool for twitter based on galois lattices”, In Conceptual Structures for Discovering Knowledge, pp. 91-103, Springer Berlin Heidelberg, 2011.

[17] M. Aufaure and B. Le Grand, “Advances in FCA-based Applications for Social Networks Analysis”, International Journal of Conceptual Structures and Smart Applications (IJCSSA) 1, vol. 1, pp. 73-89, 2013.

[18] C. Carpineto and G. Romano, “Concept data analysis: theory and applications”, Wiley, 2004.

[19] S. M. Dias and N. J. Vieria, “A framework for the development of formal concept analysis algorithms” in portuguese “Um arcabouço para desenvolvimento de algoritmos da análise formal de conceitos”, Revista de Informática Teórica e Aplicada, vol. 18, no. 1, p. 31-57, 2011.