Paper 106-2012

Community Detection to Identify Fraud Events in Telecommunications Networks

Carlos André Reis Pinheiro, Oi, Rio de Janeiro, Brazil

ABSTRACT

Telecommunications' industry evolves into a high competitive market which demands companies to establish an effective revenue assurance framework. Social network analysis can be used to increase the knowledge about the customers' behavior, not just in terms of individual usage but mostly in relation to the customers' connections and how they create communities according to their call and text messages. By performing community detection, telecommunications companies are able to recognize groups of customers which unexpected behavior in terms of usage and also in regard to types of social structures. Outliers groups might be pointed out as suspicious communities in terms of fraud events, delivering a relevant knowledge about possible leakages of money.

INTRODUCTION

This paper presents how social network analysis and community detection based on calls and text messages upon communications network might be used to understand customers' behavior in many different aspects. The analyses of social relationships may point out the distinct aspects of customer behavior and the viral effect of these events throughout the network, or within communities. This viral effect or diffusion cascade into communities may be used to explain the way some particular customers lead others in making churn or in purchasing products. Once fraud is a regular event, a business event for the fraudsters, at least, a type of influence might take place as well within the members of a community. Based on that, social structures evaluation might highlight outlier's groups of customers in terms of usage, indicating possible occurrences of fraud within the network. Social Network Analysis can deliver a relevant centric overview about the customers, including the way they use products and services, they influence others, and even the way they commit fraud.

Social network analysis can raises relevant knowledge about customers, particularly in relation to the way they relate to each other, and therefore, the process of influence within social structures. The overall method to build and analyze social networks, as well as the metrics assigned to this technique, will be described in this paper in order to support the case study presented. In this particular case, all analyses are based on communities comprised within telecommunications networks and therefore, all individual social metrics are compared against to the communities' measures they are fitted in. An outlier analysis approach is performed over those social metrics, both, considering the social measures for individuals and communities.

Most analytical approaches take into consideration usage and demographic variables, which describe the individual behavior assigned to the customers. Social network analysis on the other hand evaluates the relationships among customers, regardless their individual information. Customers are evaluated rather based on their connections than their individual attributes, which might highlight how important they are within social structures.

This study presents a significant augment in terms of fraud detection both considering the process to simple understand the fraud behavior and the predictive modeling of fraud events.

THE COMPRISED SOCIAL NETWORKS BEHIND TELECOMMUNICATIONS

The main feature of any community is the relationship between their members. Any kind of community is established and maintained based on this sort of connection. Also, the telecommunications environment is evolving into a wide scope of new devices and technologies, which spread out the possible types of communications among customers. Due to these brand new technologies and communication devices available currently, this type of characteristic have been increasing in relevance and becoming more evident in telecommunications industry. From different possibilities of communication and relationship, with flexibility and mobility, those communities gained new boundaries, creating new means of relationships and therefore into social networks. Based on different ways to communicate, people are ready to create new types of virtual networks. Due the dynamism and cohesion of these communities, included in a context highly technological, the people's influence can be significantly more strong and relevant for telecommunications companies.

Additionally, some sort of events can be understood as happing in a chain process, which means that some particular point of the network might trigger a sequence of similar events. This type of chain can be started as a matter of influence or as simply word of mouth process. As long as customers get knowledge about something, they are able to spread this knowledge throughout their social networks.

Customers might influence others to make churn, to purchase some service or to adopt a sort of product. However, customers, or fraudster in this particular case, might influence others to commit fraud. Fraud is a business, and once fraudsters establish a properly environment to make money, they spread out that environment over social structures, allowing customers to make fraud against telecommunications companies. As social structures, communities can also spread out the word of mouth about a particular type of fraud to other communities, increasing even more the loss for telecommunications companies. All these new technologies make customers to get in touch to each other even faster, and therefore, a particular fraud event can be spanned throughout social networks quite quickly.

The linkages between customers can describe the way that the events will run and the weight of the linkages can represent the impact of the chain process in the network. Understanding the way these relationships take place within the network and recognizing the influences of some specific points inside the network could be a relevant indication where fraud beginning.

Due to the almost one to one relationship between telephones and individuals, and the influence which some person could exert upon other members inside communities, it is quite important to recognize the main characteristics for such social structures and therefore to highlight suspicious activities within the network. These suspicious activities should be forwarded for further investigation, validating the fraud occurrences and thus leading the learning process assigned to the analytical model.

Discovering the central, strong, or hub nodes, for instance, it is possible to understanding how a particular event spreads out within the network, and also how customers follow that event over the time. In this way, it is possible to identify the root of the fraud event and the mobility of it throughout communities.

USING SOCIAL NETWORK ANALYSIS TO DETECT OUTLIERS

Social network analysis can disclose the possible correlations among some particular business event, such as usage or consume. Monitoring and analyzing social structures over the time, particularly those ones are interconnected, can allow companies to recognize suspicious events within communities and therefore point out possible revenue leakage.

This paper aims to present the social network analysis based on data from a telecommunications company. This data is in relation to the calls among the customers. The main objective of the social network analysis in this particular case is to highlight unusual behavior in relation to customers or communities. The idea behind of using social analysis is to identify unexpected relations assigned to individual customers, and assigned to customers considering their communities. This approach can reveal possible suspicious connections inside the network, indicating as consequence the calls which might likely to be fraudulent.

A distinguish approach to identify suspicious communities due to the network measures is presented, describing the steps in relation to the process of building the networks, computing the metrics, and finally and most important, evaluating the network analysis outcome. The methods in conjunction; based on social network analysis and exploratory analysis; can raise relevant knowledge and understanding about the fraudsters and mostly the consumers who follow them in a particular of event of fraud.

There are two lines of analyses in here. The first one is to detect the viral effect of fraud within communities. What happens onwards with the other members in a community when one particular customer commits fraud in a point of the time. Do they commit fraud afterwards? This is the viral influence of fraudsters within communities. It is the word of mouth into groups of customers when they get aware a possible fraud is coming up.

The second one is the outlier analysis of individuals within communities considering the social metrics, such as degree; in and out; influence, closeness and betweenness, hub and authority, and page rank.

The social network is established due to the call detail records, connecting the caller to the called costumers upon a link. Both nodes (customers, or telephone numbers) and links (calls) can be weighted in order to distinguish the importance of them. Important nodes contribute more in some particular network measures, as well as important links. The weight of the nodes in this specific study is based on the customers' average billing, and the weight of the links is based on the call's value, the frequency and also the duration. All calls are aggregated by type, upon a relation between the frequency and the total duration.

Once the network is built, a procedure of community detection is performed in order to identify the existing groups of customers who are related to each other. Then, an exploratory analysis is executed in order to highlight unusual or unexpected behavior for the groups of consumers or for consumers within communities. The unexpected behavior is pointed out due to the outlier observation. Distinct percentiles might be defined due to the distribution values for the network (individuals and communities) measures, as described above. This is certainly a heuristic process to find the optimal cutoff, and it is also due to the capacity to handle the alerts raised by rules according to the network and exploratory analyses.

Analogously to any traditional exploratory analysis, the presence of outliers indicates an unexpected behavior, which arise a particular set of rules and thresholds to highlight suspicious events. The main difference here is that according to the social network analysis, the occurrences of outliers are rather raised based on the relationships than to the individual attributes. Particular values assigned to groups of nodes might indicate unusual behavior, as well as, the sum of individual values for all nodes from a group.

COMMUNITY DETECTION TO DEFINE GROUPS OF USERS

One of the most important outcomes from a social network analysis is the measures in relation to consumers. How they relate to each other, in which frequency, and how relevant are their connections. From the network analysis perspective, the suspicious nodes might be pointed out by the nodes measures itself, by the links measures, or even by the communities' measures. Nodes with very high values for some network metrics might be suspicious. The nodes comprised in links with high values network metrics might be suspicious, and finally, nodes comprised in communities with high values network metrics might be also suspicious.

There are some relevant concepts in relation to the social network analysis approach, which describes the importance of some particular nodes, links, groups of nodes and the overall social structure. These concepts are used in order to highlight suspicious nodes within communities.

The method established in this case is firstly detect the communities, and then, compute the social metrics for each node considering the communities they are comprised. There are multiple possible communities to be identified, and some basic rules might be put in place in order to deal with it.

There are some techniques which are put in place in order to identify communities inside the entire network. Differently of the connected components, which are isolated from the rest of the network, a community can holds some nodes which have connections outside the community, as branches or arms outside the internal community. In this way, a community can be connected with other communities, though more than one node. Communities inside networks can be understood as clusters inside populations. The study of the communities can raise some important knowledge in relation to the network's behaviors allowing particular analyses in terms of suspicious individuals inside them.

In big networks such as in telecommunications, it is quite often to find large communities, which follows a power law distribution, a few amount of communities containing a large number of nodes and the majority of communities containing a few number of nodes.

To diminish this problem, it is possible to work with different values of resolutions, creating by them distinct distributions of communities. The resolution defines how the network is divided into communities. Resolution is a reference measure used by the algorithm to decide whether merge or not two communities according to the intercommunity links. In practice, larger resolution values produce more communities with smaller number of members within them. In order to produce big communities, the algorithm needs to consider more links among the nodes, even the weak connections which bind them. Thus, bigger communities mean larger amounts of members, and therefore, weaker strength in the links which connect them. On the other hand, smaller communities mean fewer amounts of members but holding stronger links among the nodes. There is also a metric which indicates a possible best resolution for the community detection, called modularity. As bigger the value the best is the communities' distribution in terms of number of members. However, the decision of the number of communities and the average amount of members could be made upon business needs. Fraud might require strong links among the members, as well as churn. On the other hand, purchasing and product adoption might require the higher amount of members within communities so the viral effect may be wider.

As a heuristic process, there is no formula to identify the best community´s distribution, or the best resolution number to divide the network into communities. This process is a try and error approach, according to the size of the network, the business problem to be solved and the onward analytical development to be made.

METHODOLOGY TO DETECT OUTLIERS INSIDE THE SOCIAL NETWORKS

The next phase of the network analysis approach is to consider the measures for groups in comparison to the overall values so the outliers can be recognized. This approach arise relevant aspects in relation to the entire network and mostly in respect to groups of consumers. There are several network metrics to be calculated in relation to social network analysis. In this particular approach, the social metrics described onwards are computed, always in relation to the communities found. It is important to emphasize that, first, the communities are detected and therefore the social measures are computed upon these communities.

Degree

The degree centrality represents the number of connections a particular node has. In a directed graph, where the direction of the node is relevant, especially in telecommunications, there is a differentiation between the in-degree;

the number of links a particular node receives, and the out-degree; the number of links a particular node sends. The sum of in-degree and the out-degree gives the degree measure.

Closeness

The closeness measure represents the mean of the geodesic distances (shortest path in the social network perspective) between some particular node and all other nodes connected with it. This measure describes the average distances between one node and all other nodes connected with it. It can be understand as how long a message will take to spread throughout the network from a particular node n. It is a measure which describes the speed of the message within social structures.

Betweenness

The betweenness measure represents how many shortest paths a particular node makes part. Nodes that occur on many shortest paths between other nodes have higher betweenness than those that do not. It can be understand as how central a node is considering for the entire network and all connections it has. It also represents how far a message can reach within a network from a particular node n. It is as measure which describes the span of the message within social structures.

Influence 1

The centrality measure influence 1 represents the first order centrality for a particular node, which means how many other nodes it is straight connected. This measure can be understood as how many “friends” a particular node has. It describes how many nodes can be possible straight influenced by some particular node. The influence 1 for a particular node n considers the weight of nodes and links adjacent to n.

Influence 2

The centrality measure influence 2 represents the second order centrality for a particular node, which means how many nodes the nodes it is straight connected are connected. This measure can be understood as how many “friends” my “friends” have. It describes how many nodes can be possible influenced by some particular node. The influence 2 for a particular node n considers the weights of nodes and links adjacent to n as well as the weight of the links for the nodes adjacent to the nodes adjacent to n.

Hub

The hub measure represents the number of important nodes a particular node n point to. This measure describes how this particular node refers to other important nodes. As much important nodes it refers, more important it is.

Authority

The authority measure represents the number of important nodes point to a particular node n. This measure describes how this node is referenced. As much important nodes refer to it, more important it is.

Page Rank

The page rank measure represents the percentage of possible time that other nodes within the network might spend with a particular node n. It is originally a algorithm from Google to rank websites according to the search performed. In this particular study, considering the telecommunications environment, this measure infers how possible this particular node n might be important, whether referring other nodes or being referenced.

All these measures are taken into account to highlight the outliers in terms of occurrences within the social structures of the telecommunications network.

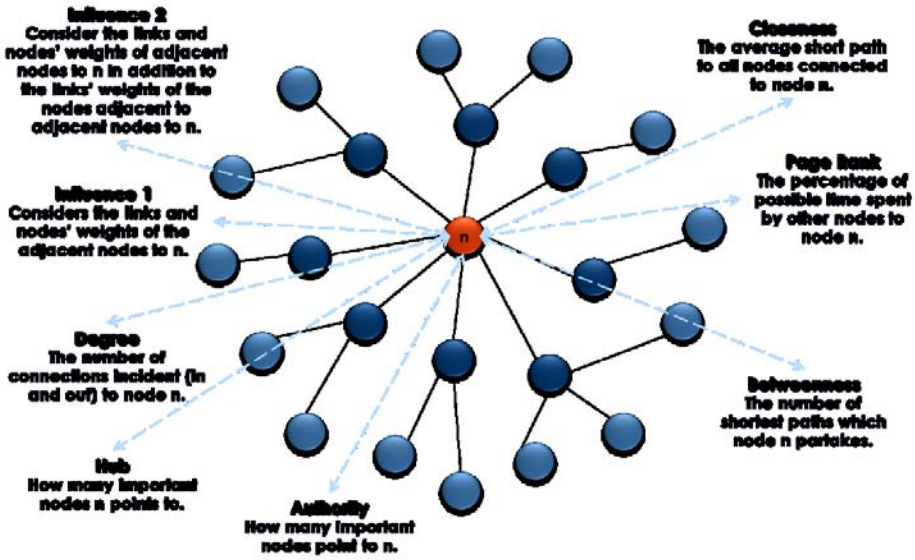

Figure 1. Social measures computed upon communities to highlight outliers nodes

EXPLORATORY ANALYSIS UPON THE SOCIAL NETWORK ANALYSIS OUTCOMES

The detection of outliers inside the network, by computing network measures for nodes and links and therefore by applying exploratory analysis is one of the possible methods to recognize unusual behavior, not in relation to individual calls and texts but in respect to the groups of consumers and their behavior in terms of usage.

Isolate calls might hold expected values for the some particular attributes such as origin, destination, durations, type, time, day, and so forth. However, the relationship among the consumers due to their calls might reveal unusual frequency, recency or density in respect to some network metrics. Social network analysis can therefore reveal unexpected behavior for nodes when they are considered as a group.

As a business, fraud events have a beginning, which is clearly identified when fraudster create the properly environment to diffuse some illegal usage of the telecommunications network. As long as this environment is ready to use, the fraud events start spanning throughout the network, diffusing within social structures. Very often, that event is highly concentrated among the small groups that are aware about the fraud, and hence, will present a quite different type of behavior than the regular groups of consumers.

The outlier analysis upon the nodes as individuals might don't indicate the suspicious behavior. Instead of that, the outlier analysis upon groups of consumers or communities can lead the investigation toward to suspicious cluster, pointing individually the nodes comprised within these groups. It is slightly different in terms of procedures but achieving totally distinct results.

RULES AND THRESHOLDS BASED ON OUTLIER ANALYSIS

After computing the social metrics for all communities, considering all social measures previously described, an outlier analysis is performed so the unusual behavior can be raised. A univariate analysis might be computed in order to classify all observations into particular percentiles and therefore separate the higher points. For instance, taking the percentile 1% or 5% it is possible to highlight X number of nodes, comprised into Y number of communities. These X number of nodes holds an average value for all social network measures, such as degree-in A, degree-out B, closeness C, betweenness D, influence 1 E, influence 2 F, hub G, authority H, and page rank I. Also, we might consider the number of members comprised in the groups identified and the density of some particular network metrics assigned to each group. By applying these thresholds (A, B, C, D, E, F, G, H, I) for the whole network, a Z number of nodes will be raised, indicating the possible observations to be sent for further investigation in terms of fraud risk or suspicious.

This approach consists of a sequence of steps, from identifying the set of communities comprised into the network, computing the social network metrics for each community, identifying the outlier percentile (1% for instance), establishing the average values for the outliers' network measures and then applying these thresholds to the entire network. The outcome of this sequence is a particular set of nodes which require additional investigation to validate the occurrence of fraud. The outlier percentile, which will determine the number of nodes raised as suspicious is a matter of the distribution values for all network metrics but also the operational capacity to handle the further

investigation by the fraud analysts. Lower cutoffs for the outlier percentile might raise a greater number of nodes, and higher cutoffs for the outlier percentile might raise a less number of nodes to be analyzed.

CONCLUSION

The following figures present some suspicious behavior when the exploratory analysis considers the social network measures. The red node in the middle of the social structure presented as follow makes calls to almost seven thousand distinct phone numbers, which, as a matter of fact, is quite suspicious. Its social measure **degree-out** is absolutely huge in comparison to the rest of the network, indicating this particular node as a outlier, and therefore, to be send it through for further investigation. The nodes it is related to might be also a target for additional investigation.

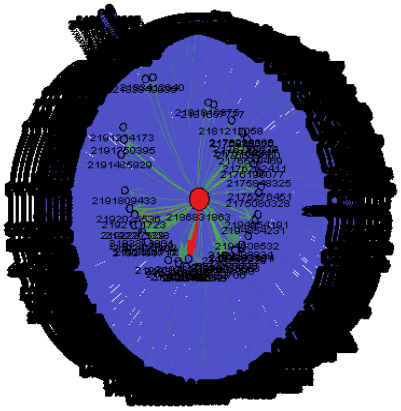

Figure 2. Outlier in the social measure degree-out

The red node in the middle of the social structure presented below receives calls from more than one thousand distinct phones, which, in analogous approach, is quite suspicious. In that particular case, the social measure **degree-in** is used to highlight the outlier observation.

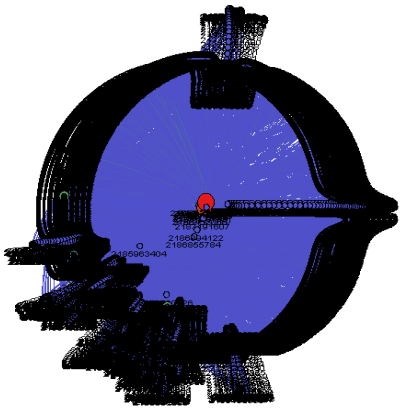

Figure 3. Outlier in the social measure degree-in

The big node in red in the middle of the social structure showed below makes more than sixteen thousand calls to the node very beside. The thin arrow represents how strong this connection is in comparison to the other links. A combination of the links strength and the social measure **hub** is used to highlight this particular outlier.

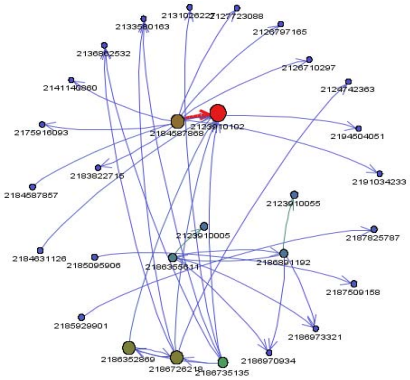

Figure 4. Outlier in the link strength plus the social measure hub

Finally, the red node with thin arrows in the social structure showed below receives more than twenty thousand calls from just twenty six distinct phone numbers, represented by the nodes surrounded it. Analogously, the thin arrows represent the strength of the links among this particular node and its correlated connections. A combination of the links strengths and the social measure **authority** is used to highlight this particular outlier.

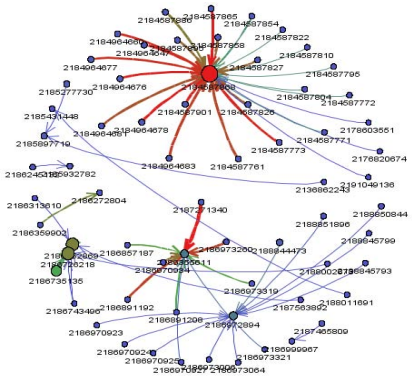

Figure 5. Outlier in the link strength plus the social measure authority

As in the most of the procedures assigned to fraud management, this approach points out suspicious events of fraud, which in practice, indicates these events for further investigation. The outlier analysis upon the social network measures describes a possible behavior for groups of consumers committing fraud. This average behavior can be translated into a set of rules and thresholds in relation to network metrics. This set of rules and thresholds might be deployed into transactional systems for instance to raise alerts about possible events of fraud. Alerts are not stamps of fraud, but instead of it, they are indications of how likely these particular events might be fraudulent.

As a matter of fact, as any other analytics initiative, the outcomes from mathematical or statistical procedures should be taken into account to make decisions, not to decide straightly. The analytics process is a support to make better decisions, not a surrogate method to the decision making procedure.

REFERENCES

Carrington, P. J., Scott, J.,Wasserman, S. 2005. *Models and Methods in Social Network Analysis*. New York: Cambridge University Press

Degenne, M. F. 1999. *Introducing Social Networks*. London: Sage Publications

Knoke, D., Yank, S. 2007. *Social Network Analysis*. London: Sage Publications

Pinheiro, Carlos Andre Reis Pinheiro, 2011. *Social Network Analysis in Telecommunications*. 304 pages. Hoboken, New Jersey: John Wiley & Sons Inc

SAS Institute Inc, 2011. *SAS 9.2 OPTGRAPH Procedure: Graph Algorithms and Network Analysis*. 198 pages. Cary, North Caroline: SAS Institute Inc

Wasserman, S., Faust, K. 1994. *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press

ACKNOWLEDGMENTS

Oi and SAS Brazil were crucial to develop and deploy the analytical model present in this case study. The partnership with SAS dramatically improved the outcomes of this project.

RECOMMENDED READING

- Social Network Analysis in Telecommunications, Wiley and SAS Business Series

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Carlos Andre Reis Pinheiro
Enterprise: Oi
E-mail: carlos.pinheiro@oi.net.br