USING ROOT MEAN SQAURE FOR LONGITUDINAL APPROACHES IN SOCIAL NETWORKS

Rasim Alguliyev[1], Yadigar Imamverdiyev[1], Hamid Zargari Asl[1-3], Maryam Bazel[2]

[1]Institute of Information Technology of ANAS, Baku, Azerbaijan
[2]Islamic Azad University, Ardabil, Iran
[3]Mobile Communication Iran, Ardabil, Iran
[1]*yadigar@lan.ab.az*, [2]*hamid_zargari@ardabiltelecom.ir*

Abstract

Longitudinal studying the networks is repeatedly measurement of networks on a given node set in a time period. It is also related to Social Network Change Detection (SNCD) that is used in detection the occurred changes in a given network. In this study we offer a new method for finding a reasonable threshold for indication the border of normal and abnormal changes (decision interval) in one edge of a social graph by using an electrical concept. It can be obviously applied entirely to the whole network.

Introduction

The social network is investigated in different groups and media and used to recognize and distinguish more strong or more week ties between members. The results are offered to assist the management to choice the suitable medium to a specific group and also use to assess the influence of the problems of connectivity in disrupt the ties or upgrade the level of mutual exchanges [1].

As a computing platform, mobile phones are both secret and personal. Individuals store private information on them and often personalize their appearance or ring tones, for example. Smart phones are a particularly tempting platform for building context-aware applications because they're programmable and often use well-known operating systems [2]. Continuous behavioral data logged by the mobile phones compared with self-report relational data yields the information from these two data sources. A new method presented for precise measurements of large-scale human behavior based on contextualized proximity and communication data alone, and identifies characteristic behavioral signatures of relationships that allowed us to accurately predict 95% of the reciprocated friendships in the study [3]. A lot of studies have been achieved to realize the patterns and use them for discovery the subgroups or missed links and they have offered different methods and resorts but the noises finding that is used to hiding the illicit activities is worth investigating. By magnifying the small abnormalities we can achieve a lot of fruitful targets [4].

Longitudinal Concept and Simulator

We focus on social graphs whose members are dedicated to every connection between an actor and its counterpart. On the other hand the number of the connections between actors in a time interval may be illustrated by weight of edge. We suppose that matrices are arranged by their time indices i.e. every matrix is dedicated to a particular time period. The latter 2D matrices can be supposed as following matrices: A_{t0}, A_{t1}, A_{t2}, …

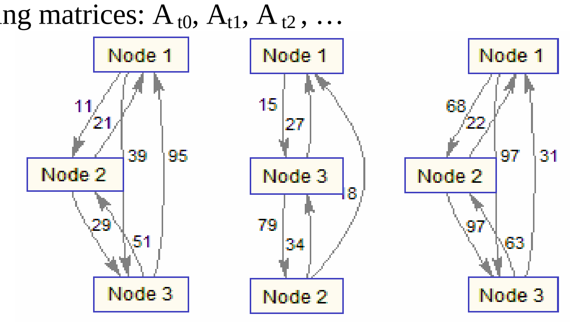

Figure 1. Three actors in social graph in 3 sampling time periods.

By following pseudo code we make an independent array which demonstrates the variation of weights in every link between two actors. The third dimension is showing the alteration of every edge during the period of sampling. Figure 2 shows the way that is used to make new array.

```
for (i=0;i<MAX_ACTOR_No;i++)
                        /* traversing the lines */
    for (j=0;j< MAX_ACTOR_No;j++)
                            /* traversing the columns */
        for(k=0;k<MAX_TIME_PERIOD_No;k++)   NewArray[i][j][k] =A_tk [i][j];
                                /* making 3D array comprising all 2D matrices*/
```

For evaluation of our claims we made a simulator which is producing social matrices by randomly generated matrices. The next pseudo code illustrates the simulator functionality.

```
for(i=0;i<MAX_ACTOR_No;I++)
            /* traversing the lines */
    for (j=0;j< MAX_ACTOR_No;j++)
                /* traversing the columns */
            for (k=0;k<MAX_TIME_PERIOD_No;K:)
                    /* traversing time vector */
                NewArray[i][j][k]=random( MAX_WEIGHT-BOUND);
                        /* dedication the random values */
```

Concentrating on one member leads us to draw the curve of weights. This is supposed as *V(t)* similar to a voltage function in electrical engineering and it is feasible to prepare integral, derivation and so on as well. Regarding to normal connections those are continually setting up, they can be cancelled by making the first derivation which is illustration the changes, thought in some works Fourier transform is used [5]. It is obviously realized that normal activities will have small deviations from a specific range and definition the upper and lower threshold can be important criteria for classification the activities. Once passing beyond the border (decision interval) will announce us that there is a violation the normal behavior. In figure 2 the method of preparation the curve is depicted.

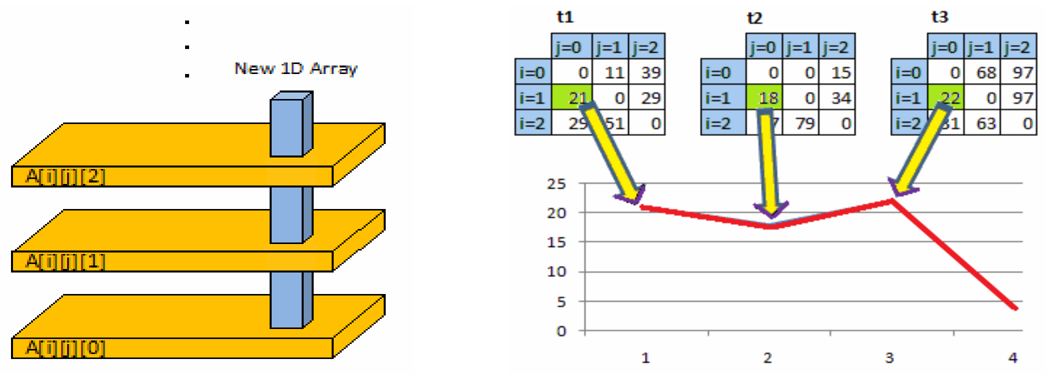

Figure 2. Selection only one member of 3D array (left) and preparation its cureve (right).

Figure 2 (right) shows a curve that is the consequence of one sequence of numbers (weight of edges) and it is possible to make the first derivation by using *FirstDifferential[i][j][m]=(NewArray[i][j][k+1]-NewArray[i][j][k])/TIME_PERIOD* statement. The result array will contain MAX_TIME_PERIOD_NUMBER-1 elements. In turn the second array therefore can be manipulated in the same way. By application this algorithm in our numbers and $t_{k+1} - t_k = 1$ assumption the following arrays will be yielded:

FirstDifferential[i][j][m]'s elements: 15, -33, 40

The latter numbers can be used for mining statistical parameters such as average, standard deviation and so on which can be used for finding our thresholds to purge the unnecessary information.

In this approach we handled one hundred 3*3 tables made by random numbers which ranged from 50 to 110 for every cell. Furthermore we classified the cell's contents with phrase:

*NewArray[i][j][k]=((i+1)*30+(j+1)*10)*random(1)*

This pseudo code changes the variation of every cell so that range of generated random numbers in every cell differs from others and it is ascending regarding lines and columns. The last formula gives the coefficients of random numbers and the result factors will be the following set: 40, 50, 60, 70, 80, 90, 100, 110, 120. It should be mentioned that diagonal cells of matrices must be zero. This categorization will help us in further analysis and evaluation our algorithm.

This simulator that we created is generating automatically 100 matrices and then makes 99 matrices which denote to the 99 time intervals which are supposed between every sampling period. This is first derivation of links. We compared the results in two members of matrices called *NewArray[0][1]* and *NewArray[2][1]* because of their obvious differences in ranges that are dedicated to generate random numbers.

Figure 3 (right) shows the derivation of curve that has been illustrated in left part of figure 3. The simulator makes the root mean square (RMS) numbers automatically step by step and by making "square", "average" and finally "root" component. Table 1 and table 2 show the sample output of simulator. It is possible to draw a boundary line (figure 4) using the delivered number which implies to the threshold of observed abnormal time periods in edge's life time (The curve in figure 4 is made by another random number set).

In related works it has been called as decision interval, where the threshold is applied by analyst and on betweenness network centralization [5].

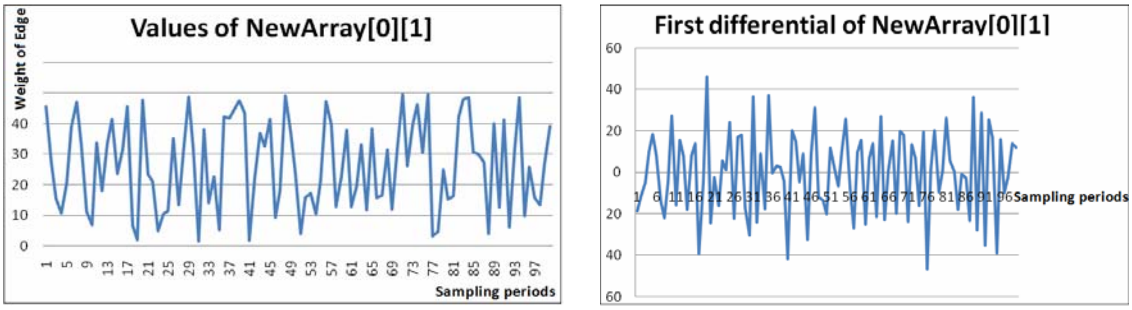

Figure 3. Sample curve made by simulator (left) and derivation curve (right).

Table 1. Generated Random Matrices

		Generated Matrices in 100 supposed time periods:										
	Matrix		1	2	3	4	5	6	7	8	9	10
f(x)	sery 0,0		0	0	0	0	0	0	0	0	0	0
	sery 0,1		0.36645	4.752864	37.78838	45.70008	15.36286	20.28362	43.3236	11.12551	36.872	19.43167
	sery 0,2		0.061271	30.652	42.63781	5.702158	0.737397	17.09279	41.50131	21.74172	10.40976	49.22239
	sery 1,0		50.89844	68.8739	19.0548	32.76668	8.076454	55.90405	55.66798	69.76883	37.07311	19.34365
	sery 1,1		0	0	0	0	0	0	0	0	0	0
	sery 1,2		82.01463	52.17285	40.60279	88.42211	49.98159	3.331875	28.23423	76.99679	64.41938	46.58241
	sery 2,0		5.885065	17.46081	22.81395	9.972183	62.91023	30.58989	18.4763	74.77463	78.20709	35.43642
	sery 2,1		24.13633	45.49886	62.18017	76.46634	20.58908	30.88067	91.75232	109.6177	47.44938	14.21465
	sery 2,2		0	0	0	0	0	0	0	0	0	0
f'(x)	sery 0,0		0	0	0	0	0	0	0	0	0	0
	sery 0,1		4.386414	33.03552	7.911691	-30.3372	4.920761	23.03998	-32.1981	25.74649	-17.4403	-12.609
	sery 0,2		30.59073	11.98581	-36.9357	-4.96476	16.35539	24.40852	-19.7596	-11.332	38.81263	-38.0486
	sery 1,0		17.97546	-49.8191	13.71187	-24.6902	47.8276	-0.23607	14.10084	-32.6957	-17.7295	7.660702
	sery 1,1		0	0	0	0	0	0	0	0	0	0
	sery 1,2		-29.8418	-11.5701	47.81931	-38.4405	-46.6497	24.90236	48.76256	-12.5774	-17.837	-27.3978
	sery 2,0		11.57574	5.353141	-12.8418	52.93805	-32.3203	-12.1136	56.29833	3.432462	-42.7707	-19.1896
	sery 2,1		21.36253	16.68131	14.28617	-55.8773	10.29158	60.87166	17.86535	-62.1683	-33.2347	57.58935
	sery 2,2		0	0	0	0	0	0	0	0	0	0

Table 2. Root Mean Square Calculation

Square:	8.214776	17.29092	79.08924	300.743	944.7991
Average:	400.7972				
Root:	20.01992				

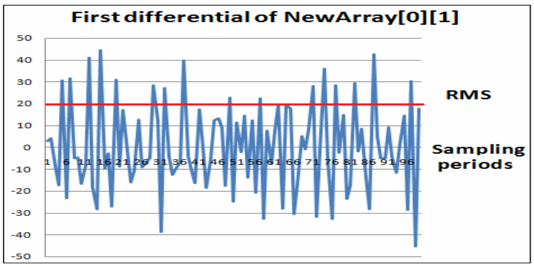

Figure 4 Derivation with RMS line (RED).

Results

In every mobile communication office the data about calling and called subscribers are stored in Call Detail Records (CDR) and they are available to be used as log files [6]. We focused on the collected log of two actors in a social network as two employees and used their information to assess our algorithm. The time of established calls in a time period is supposed as weights of edges and the time in every call in the same day were added to the total time. The obtained results comprising original values, derivation and RMS is shown in figure 5. The abnormal points in 5th and 22nd days of first month and in 17th day of second month were compared with self reported information of volunteer actor. Comparing the activities of focused actor illustrates that in those days he was in journey and out of his homeland. This is very desirably accommodated with our predefined algorithm.

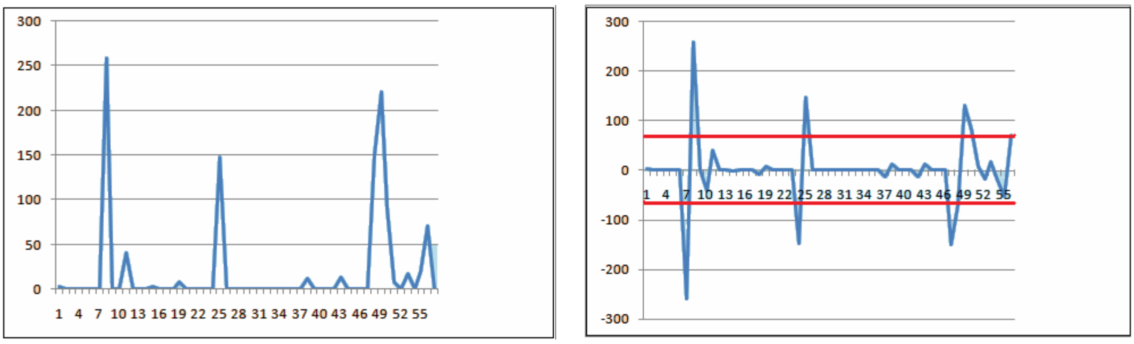
Figure 5 Curve of weights (left) and derivation curve with RMS red lines (right).

The activity curve of one specific connection that can be traced is changed to differential form (derivation), this is a method to show how does the relation ascend or descend. We worked on one pair of actors but it can be used for a social network. Preparation the RMS which is used in electrical engineering industry is acceptable as a way for finding the stable value of a fluctuating curve. It can be also stated that it illustrates the alteration of connection or weight of edges which is one part of longitudinal studies.

References

1. Caroline Haythornthwaite, "Social Networks and Internet Connectivity Effects." June issue of the journal Information, communication & Society Vol8, No2, 2005, pp.125-147.
2. Mika Raento, Antti Oulasvirta, Renaud Petit, Hannu Toivonen, "Contextphone: A prototyping platform for context-aware mobile applications", Published by the IEEE CS and IEEE ComSoc PERVASIVEcomputing 51, 2005.
3. Nathan Eagle, Alex (Sandy) Pentland, David Lazer, "Inferring social network structure using mobile phone data", MIT Design Laboratory, Massachusetts Institute of Technology, Cambridge, MA, PNAS http:/www.media mit.edu.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 2007.
4. Jafar Adibi, Paul R. Cohen, Clayton T. Morrison. "Measuring Confidence Intervals in Link Discovery: A Bootstrap Approach" ICDM-04: IEEE International Conference on Data Mining, 2004.
5. Ian McCulloh & Kathleen Carley, "Longitudinal Dynamic Network Analysis -Using the Over Time Viewer Feature in ORA", Institute for Software Research, School of Computer Science ,Carnegie Mellon University, Pittsburgh, PA 15213, March 9, 2009
6. International Telecommunication Union, www.itu.int