Using Rate Equation for Modeling Triad Dynamics on Instagram

Mariusz Kamola*†
Institute of Control and Computation Engineering
Warsaw University of Technology
ul. Nowowiejska 15/19, 00-65 Warsaw, Poland
Email: *M.Kamola@ia.pw.edu.pl

Research and Academic Computer Network
ul. Kolska 12, 01-045 Warsaw, Poland
Email: †Mariusz.Kamola@nask.pl

***Abstract*—Triadic analysis is a convenient way to assess structure and stability of a graph. This paper verifies stability of one triad type that is commonly perceived as transient and unstable, on Instagram subgraph, reconstructed by a specific crawling algorithm. Dynamics of that triad transition has been examined, wrt. degrees of that triad nodes and other basic structural properties of the triad neighborhood. Results show that the triad transition can be modeled as rate equation of order 2, and that it is fairly independent of any of the considered factors. A complete model of dynamics of triads can be further used in graph evolution forecasting, including phenomena like opinion spread or group forming, which may, in turn, affect real-life situations.**

***Index Terms*—Social network, Instagram, Graph dynamics.**

I. Introduction

Complex, networked systems, originally determined layout and operation of physical, chemical and biological processes. With the advance of civilization, they started to describe human technological and societal phenomena — with or without intention or conscience of the human itself. Currently, modern technologies made it possible to entangle people in various networks, which influence more than ever, their decisions.

The quality and strength of interpersonal forces in such networks may be represented synthetically by a set of network-wide parameters: the diameter, clustering coefficient, power law coefficient, assortativity coefficient and alike. Instead, node-specific description includes concepts of centrality, bridges etc. These terms refer to and are computed for a fixed network structure, but in reality most of them is associated with that network's dynamics: flows, spin, structure evolution. For instance, power law coefficient is closely related to probability distribution in preferential attachment network generation algorithm [1]. Eigenvalue centrality, in turn, relates to sojourn probability of a random walk process in the network.

Triadic analysis, introduced in [2], is a recognized and convenient way to describe a network at an intermediate level between the two above ones — and taking into account the interplay in smallest network communities: the triads. The frequency of triads of certain types relates closely to global network parameters. For example, frequency of transitive triads (cf. Fig. 1) can be used to calculate clustering coefficient.

Triad census is a recognized way to describe a graph, easy to interpret by social scientists. Although the analysis of more complex subgraphs has been proposed and justified [3], triadic census still remains the preferred approach to describe a complex network. Scientists from different domains consider preeminence of certain triads in the census to be the indicator of some phenomena. For instance, as regards social animals [4], triads of type 300 (`t300`) relate to chimp grooming and fight as well as to non-antagonistic relations of colobi, and to baboon fights. Instead, presence of `t030T`, `t030C` and `t021C` is characteristic in communities of macaques. Among many other examples of the utility of triad census, by the same author [5], one can find out that it may be used to detect suicidal thoughts [6]. Triadic census also turns out to be useful for analysis of collaboration networks [7], revealing hierarchical patterns (`t201`) or information flow deficiencies (`t030C`). Studies [8] show that certain classes of networks are characterized a by unique proportion of triads, regardless of the instance of the network itself. The presented, distinct classes are: word collocations in languages (irrespective of the language), social interactions, biological networks. The information distinguishing network classes is, above others, the ratio of transitive/intransitive triads, and the fraction of forbidden triads.

The fundamental rule on triad dynamics in the network was stated in [9]: 'open' triads, i.e. containing a node pair not connected in any way, do not exist long in an undirected network with edges representing true and strong emotional relationships between individuals. Such triads either develop the missing link, or destroy an existing one. Therefore, `t201` is forbidden, and `t300` — closed. The natural mutual curiosity is the basic motivation for people to complete any forbidden triad. If the missing link does not develop with time, then the common friend feels discomfort and tends to break one of his links with the non-cooperating friends. Such scenario

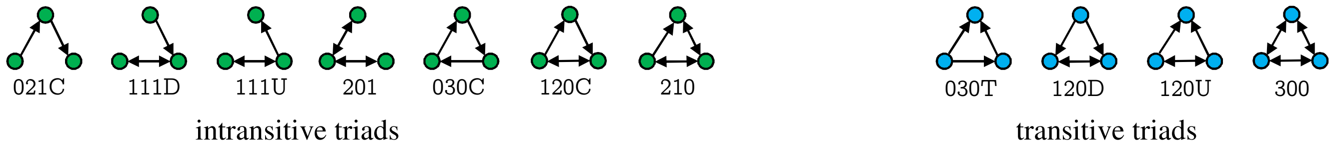

Fig. 1. Triad type notation as proposed in [2] (only consistent triads shown). Subsequent digits in triad symbols denote the number of symmetric, asymmetric and nonexistent links. Optionally, a letter disambiguates triad structure (Up, Down, Cyclic). A triad is transitive if A→B, B→C and A→C, where A, B, C are the triad nodes.

applies when the relations are authentically symmetric; a counterexample is a real estate agent and his/her customers: the links stay formally symmetric, but they are not accompanied by true information symmetry. From now on we assume that observed graph structure represents authentic relationships, and that nobody dominates on a symmetric link.

Verification of the cited rule [9] was the main goal of our research. In Sec. II we present the motivation, the choice of data source, and the mechanisms for data acquisition. In Sec. III the results concerning triad dynamics analysis are presented.

II. The Dataset

Our aim was to examine the dynamics of interpersonal relationships using a completely new dataset made out of publicly available Instagram user profiles. Instagram is a relatively new social site, still smaller in terms of the number of active users than Facebook, YouTube, Google+ and Twitter, and similar to LinkedIn and Pinterest. Most of its users are young, focused on visual communication with their peers. No forcible merge of accounts was done after Instagram has been bought by Facebook, which allowed the dynamics of social links to stay unbiased.

As we were interested in triad dynamics related to structural properties of the network, and not the content of user profiles, only the information about user A following posts by user B was scanned on Instagram website. This also helps to protect user privacy shall the dataset be made publicly available, as is currently planned. Since the download and analysis of the whole Instagram user graph was technically infeasible and unnecessary, an approach was proposed to determine, download and then watch a subgraph as dense as possible. The algorithm, shown below, performs a modified breath-first search of the whole graph, where the nodes to be visited are those having most connections with the nodes already scanned:

Require: $G(V, E), V_0, N_{\max}, t_{\max}$
Ensure: $G_0(V_0, E_0)$
  $T \leftarrow (V_0)$
  repeat
    for all $V_i \in T$ **do**
      if V_i is not an *outlier* **then**
        $V_0 \leftarrow V_0 \cup \{$all nodes followed by $V_i\}$
        $E_0 \leftarrow E_0 \cup \{$all edges from $V_i\}$
      end if
    end for
    $T \leftarrow$ nodes followed by V_0, sorted by indegree, at most $t_{\max}$ nodes
  until $\mathbf{card}(V_0) < N_{\max}$

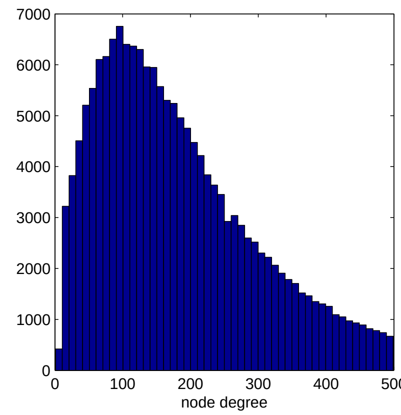

Fig. 2. Node degree distribution for the collected subgraph G_0.

Here $G(V, E)$ is the whole Instagram graph, and G_0 — its final subgraph, being further subject to monitoring. Here, the outliers are both nodes with exceptionally high indegree, especially if not balanced by high outdegree, as well as nodes not following anyone. Thus, the subgraph G_0 gets more dense in the course of the algorithm run. By its nature, the algorithm tends to include highly connected nodes, which means that finally G_0 will not be scale free cf. Fig. 2. We have get rid of extremely high degree nodes by means of outlier detection (max. indegree set to 2000, max. outdegree set to 500), and of the bulk of low degree nodes by means of the algorithm preference for highly connected individuals. Parameter $t_{\max}$ determines exploratory character of the algorithm; seeting it to a low value (4, in our case) results in stronger preference for making dense subgraphs. While this surely biases the triadic census itself, it does not impact triad dynamics in any other way that just by the number of observed cases.

The subgraph node set V_0, once determined, remained unchanged during the whole experiment. Instead, edges, or 'follows' relations, were monitored roughly every ten days, from November 2015 until March 2016, resulting in twelve edge snapshots. The number of nodes V_0 was set to 82,000, which meant that about 24 hours were required to download every single snapshot. By the end of the experiment, the dataset contained over 2 million edges, and the user locations

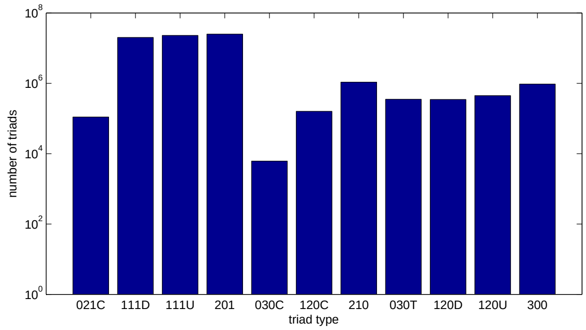

Fig. 3. Triad census for the final snapshot.

(if provided in their profiles) were mostly clustered geographically around the location of the initial root node, which is a remarkable property of the presented subgraph selection algorithm. The algorithm was not designed to record a link disruption, but severing relations with friends on Instagram was found much more rare than making a profile private or deleting it altogether.

The final triad census for the subgraph is provided in Fig. 3. Note that, contrary to the classical opinion [10], triads recognized as unstable, are well represented; in fact, they are in the majority in our graph. We may comment on that, citing [11], that social connections more often than not are asymmetric, with little authentic reciprocity. And they stay so, despite Instagram service insisting on people being followed to follow back their audience.

III. Results

A. Overall transition statistics

Fig. 4 presents all transitions between triads observed over the experiment duration. For certain triad types, there are dominant transitions. Conversions `t012D`→`t111U` and `t021U`→`t111D` mean that following or being followed by two individuals results more often in symmetry of one of the links than establishing a missing link between followers/followees (→`t030T`). Similarly, open triads `t111U` and `t111D` tend to develop symmetric links instead of the missing ones. Such effect may be a natural phenomenon or the result of Instagram incentives.

The first symptom of diversified triad dynamics is represented by triad transitions drawn with perpendicular lines. Those represent changes so rapid that they could not be recorded with our temporal scanning resolution. Some of the intermediate triads existed less than a week . They do not account for prevailing part of any of the transitions, however eg. `t102`→`t201` (30% of all transition of `t102`) is rapid. Comparing with `t201`→`t300`, we can may conclude that existence of an open triangle (the latter case) does not accelerate the establishment of a new, symmetric link a lot (38%). Later, we will examine other factors that may affect this process.

B. Triadic closure as a process

Let us focus now on dynamics analysis of the `t210`→`t300` conversion. We have chosen this case because, of all possible transitions, an asymmetric link B→C in a `t210` triad makes the highest pressure on user C to follow B. Hence, we can expect the process of establishing C→B link to be rapid. Also, limited number of possible ways the `t300` can be created makes the analysis easier.

To exclude the influence of a `t210` history, especially the triad age, we look only at triads that emerged between snapshots n and $n+1$ — therefore, examination of the number of triads that evolved into `t300` starts at $n+2$. Given the number of recorded snapshots, we have found the number of `t210`→`t300` conversions in periods between snapshots 3,4,..,8 (for `t210` triads that emerged between period 1 and 2); we repeated this procedure for two consecutive snapshots, obtaining three time series of `t210` decay rates. Those time series are presented using different colors and on various scales ways in Fig. 5.

We can perceive easily that the conversion rate decreases with time. Regardless of the reasons, such process resembles chemical or physical phenomena, which are commonly described by a general rate equation:

$$-\frac{dA}{dt} = kA^x \ , \qquad (1)$$

where A denotes state variable of the process, eg. concentration of a reactant in a chemical reaction. Depending on the exponent x value, (1) can model zero-, first- and higher order equations. To determine best matching order of the equation to the observed process, we displayed the triad conversion rates in three different ways in Fig. 5b–d. By observation, we may point out that rate vs. time graph in Fig. 5d can be approximated best by a straight line, which, despite high variability, indicates that the triad conversion process is of the second order. This has been confirmed by linear regression, which gave the MRCD (mean relative change and difference) errors of 2.64 and 0.17 for time series in Fig. 5c and d, respectively.

The graphs in Fig. 5 are non-monotonic because of some network-wide anomalies that cause user activity to be above or below modeled levels. They are shifted in the graphs because the graphs themselves refer to different absolute time. The model output, when compared with observed decay rates, as in where the ratio of those values is shown in absolute time since the experiment start, shows clearly network hyperactivity around ‘day 70’ (Jan 30, 2015). If we want to explain such global anomaly adequately, that date was also the beginning of winter school break, as well as the end of winter semester at universities in the area of our interest... As regards modeling inadequacy in days 20-30, this may be attributed to small amount of modeling data (only two scenarios available for that period). Such a graph, when based on more, filtered data, can be considered the trend of overall network activity, and used to normalize any specific time series in the future, or to model global periodic network activities.

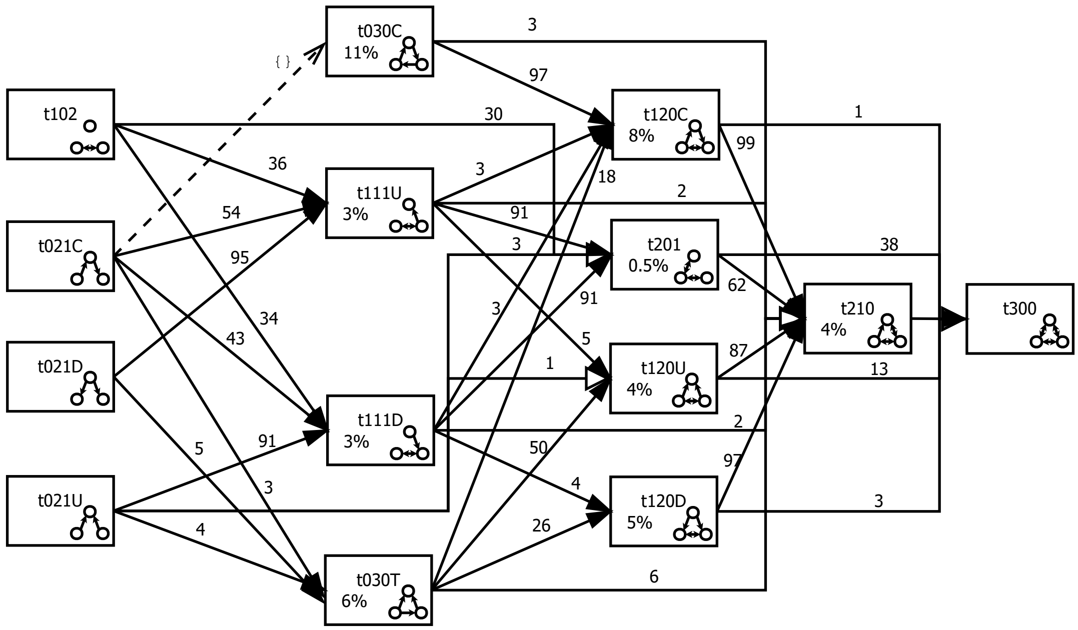

Fig. 4. Triad transition graph, leading to the final, fully closed `t300` (link disruptions are not considered here). Percentages in boxes denote the fraction of triads that have undergone any transition within timespan of the experiment. Percentages at outgoing edges denote distribution of transitions that happened.

C. Closure process parameters

Since the rate of `t210` decay into `t300` may be considered a 2-nd order equation, we may ask, what the equation coefficients depend upon? Apparently, when `t210` nodes B and C (connected still by an asymmetric link) constitute more than one `t210` triad — i.e., have more than one node A in common — they experience more pressure to develop reciprocal relation. Let us verify such hypothesis by splitting decaying triads wrt. the number of A's that B and C have in common, and identify the 2-nd order model. This is shown in Fig. 7. Interestingly, the rate of decay in case when B and C had exactly two A's in common decreases quicker that for two other segments — but this anomaly may as well be due to sparsity of modeling data, numerical errors on modeling interval boundaries, or the applied criterion for outliers filtering (those are drawn in gray color). The results, inconclusive as they are, indicate that the 'state' determining triad dynamics is unrelated to social pressure executed by common friends of nodes B and C, which conforms with earlier comments on Fig. 4.

Alternatively, we may state a more general hypothesis: if former common participation of B and C in `t201` specifically is of no importance for the dynamics, is any parameter related to B's and C's connectedness determinant? Consider the minimum of B's and C' in- and outdegrees as such a parameter of a triad, and perform modeling parameters fit, using integrated version of (1) for $x = 2$:

$$\frac{1}{A} = \frac{1}{A_0} + kt \ . \qquad (2)$$

TABLE I
2-ND ORDER EQUATION PARAMETERS, WRT. TRIAD CONNECTEDNESS

B and C connectedness	A_0 (see (2))	k (see (2))
160+	13.8	0.38
120	9.6	0.49
80	9.1	0.33
60	10.3	0.44
40	8.3	0.18

The estimated model parameters are given in Table I. As a rule, higher initial rates (A_0) must be accompanied by more dynamic decrease in time (k). However, again, the correlation between connectedness and, at least, A_0 gets disturbed by the row '60'. Otherwise, we could claim that the more connected a triad is the shorter it live before conversion into `t300` — but it would be a subtle rule, anyhow.

IV. Conclusion

The presented work aimed to detect factors that influence the process of triad transitions. Starting from the analysis of transition probabilities and distributions of transitions between possible adjacent triad types, we noticed that open triads (i.e. with missing link B-C) with asymmetric relations tend first to remove the asymmetry than the openness — cf. Sec. III-A. However, observed rapid transitions over more than one triad type indicate that triad dynamics happens on various timescales. Consequently, transition `t210`→`t300` has been examined in detail —- first, in order to determine which class

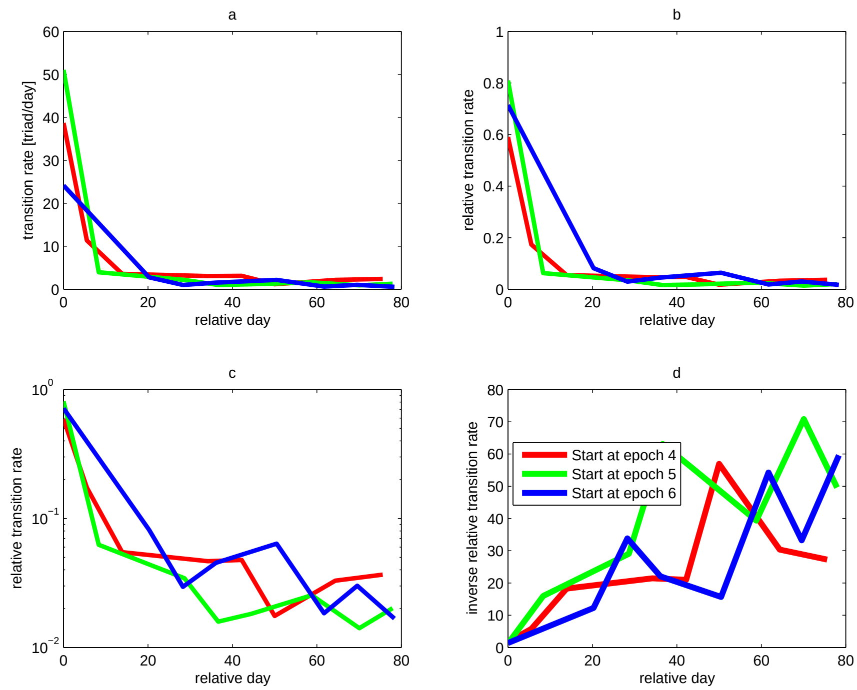

Fig. 5. Rate of new `t210` triads transition into `t300` triads, in consecutive periods since triad detection: a) number of `t210`→`t300` transitions per day; b) conversion rates normalized to the mean value; c) normalized conversion rates in logscale; d) inverse normalized conversion rates. Colors denote conversion decay process for new `t210` triads detected at snapshot 3, 4 and 5.

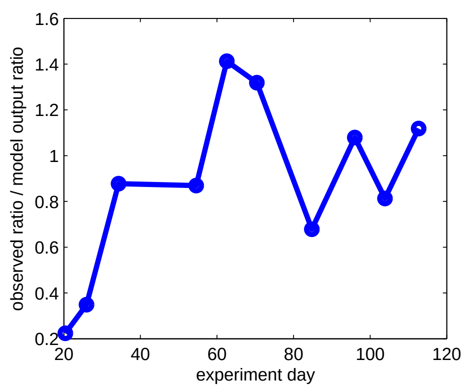

Fig. 6. The model performance in experiment timespan. Level 1 means no modeling error; more that 1 — the observed rates are higher than modeled; less than 1 — the observed rates are lower than modeled. The graph can be considered the global trend of activities in the whole network.

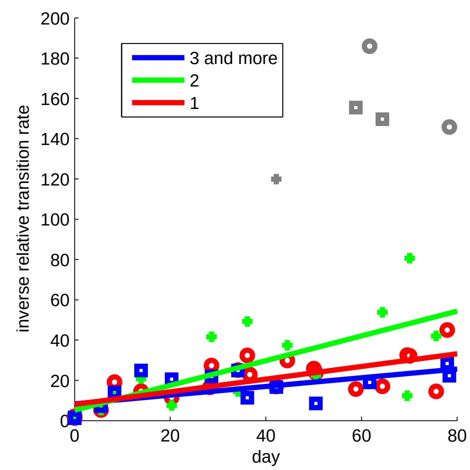

Fig. 7. Regression for coefficient k, for `t210`→`t300` decay process, split by the number of `t210` triads that contained nodes B and C (the ones that are forming a symmetric link.

of dynamic equation represents the transition rate best. Second order equation was considered the most appropriate here, as a special case of general power law approach for modeling dynamics (coincidentally with power law node degree distribution found commonly in complex social networks).

Detailed examination of influence of basic network properties of node C connecting to node B to form `t300` did not reveal any clear-cut rule. Apparently, triadic transition process is driven by factors unobservable in the dataset in possession. Alternatively, observed irregularities may derive from the modeling approach or the numerics. This should be definitely verified. Moreover, taking into account parameters describing more complex attributes than just node degree or co-appearance in multiple triads, may finally reveal more obvious rules. This has been postulated in [12] and successfully applied in practice eg. in churn prediction [13], where considering local communities greatly improved churn prediction.

The modeling efforts presented in Sec. III-B and III-C concerned only triads that *did* experience a transition — while vast majority of them stayed unchanged over the whole experiment: 120 days. So, the important unanswered question persists: what actually makes a triad prone to transition at all? In the course of experiments, we performed a number of comparisons between unchanging and changing triads, examining the same triad attributes as already presented. However, no distinction between the two classes was visible.

The presented work, apart from observation on dynamics, resulted in two valuable by-products: the algorithm for dense subgraph construction, and the Instagram dataset itself. The dataset can still be completed with unchanging extra user information, like sex, location, age or description, which may become helpful in correct segmentation of triads as regards their dynamics.

ACKNOWLEDGMENT

We would like to thank Mr. J. Jarzynski, MSc. for his commitment in the subgraph construction algorithm presented here, and in data collection activities. The results have been presented here with his kind consent, where appropriate.

REFERENCES

[1] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[2] S. L. Paul W. Holland, "Local structure in social networks," *Sociological Methodology*, vol. 7, pp. 1–45, 1976. [Online]. Available: http://www.jstor.org/stable/270703

[3] O. Frank and D. Strauss, "Markov graphs," *Journal of the american Statistical association*, vol. 81, no. 395, pp. 832–842, 1986.

[4] K. Faust, "Comparing social networks: size, density, and local structure," *Metodoloski zvezki*, vol. 3, no. 2, p. 185, 2006.

[5] ——, "Triadic configurations in limited choice sociometric networks: Empirical and theoretical results," *Social Networks*, vol. 30, no. 4, pp. 273–282, 2008.

[6] P. S. Bearman and J. Moody, "Suicide and friendships among american adolescents," *American journal of public health*, vol. 94, no. 1, pp. 89–95, 2004.

[7] A. V. Kolpakov and M. McGuire, "What can triadic census tell us about the development of public management networks?" 2010.

[8] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, vol. 303, no. 5663, pp. 1538–1542, 2004.

[9] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, pp. 1360–1380, 1973.

[10] J. A. Davis and S. Leinhardt, "The structure of positive interpersonal relations in small groups." 1967.

[11] N. Ali-Hasan and L. A. Adamic, "Expressing social relationships on the blog through links and comments," *Ann Arbor*, vol. 1001, p. 48109, 2007.

[12] T. Snijders, "Transitivity and triads," *University of Oxford. Online: http:// www. stats. ox. ac. uk/ snijders/ Trans_ Triads_ ha. pdf (last access 31-May-2016)*, 2012.

[13] W. Gruszczynski and P. Arabas, "Application of social network inferred data to churn modeling in telecoms," *Journal of Telecommunications and Information Technology*, 2016, in press.