From relation to interactions: a case study in Reddit website

Robert Jankowski[1], Piotr Górski[1]

[1]Faculty of Physics, Warsaw University of Technology

Introduction

Elucidating what factors are salient in emerging interactions in social networks is still an open question. Hence, we develop an agent-based models (ABMs) for generating interactions in signed networks. The ABMs, based on the Activity Driven Network model, use signed relations between agents to reproduce their interaction frequencies and crucial network distributions. They also consider the agents' preferences and information about neighbours of neighbours in the past. The calibration and validation step is performed on the Reddit Hyperlink Network, where agents are represented by subreddits, and links by hyperlinks between communities. We devise a profound methodology using Anderson-Darling statistic to assess the performance of the models. The proposed ABM successfully reproduces basic node-link such as a degree distribution or a link-weight distribution of the empirical dataset. Furthermore, ABMs recover a high clustering coefficient which is visible in Reddit Hyperlink Network.

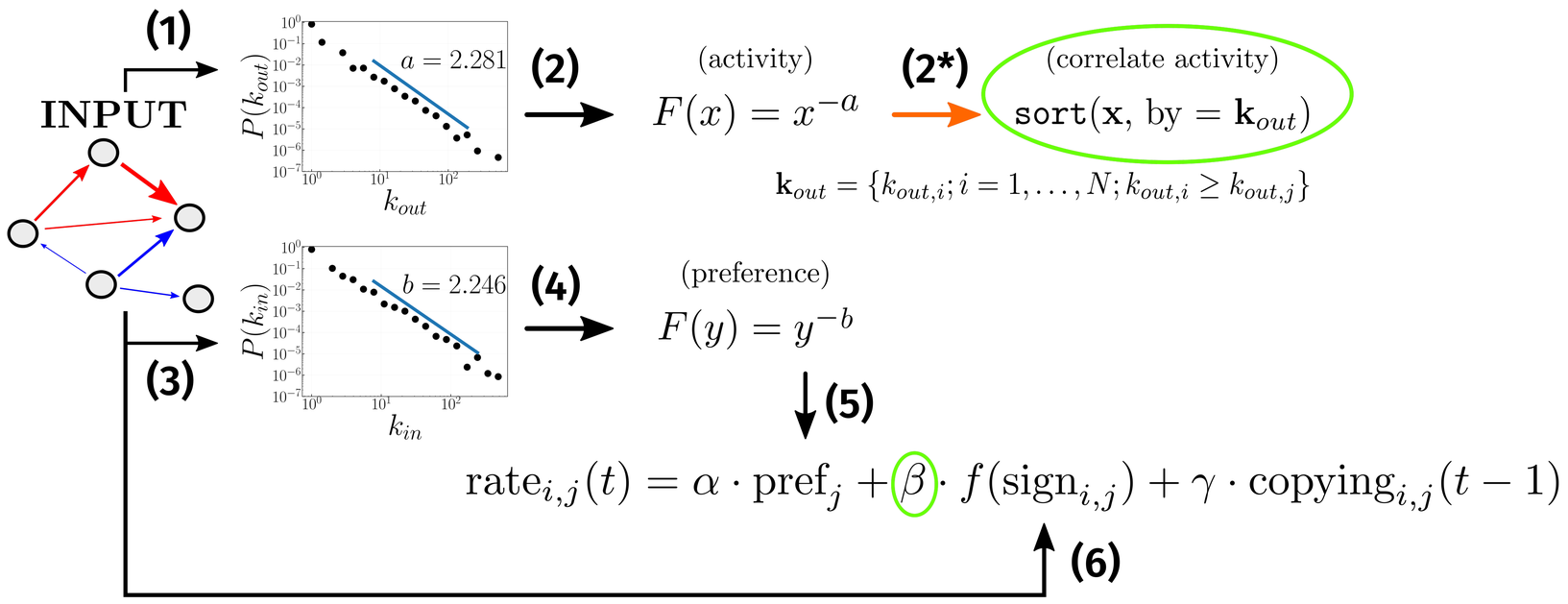

Fig. 3: A figure presents our approach to leverage signed network as an input to the ADN model. The procedure yields both activity and preference distributions from the signed input network for modified ADN model. We begin with (1) finding exponent for $P(k_{out})$ distribution of signed network, which is used in the next step (2) to generate activity distribution. Analogously, the exponent in preference distribution is obtained from $P(k_{in})$ distribution (3). Then we generate agents' preferences (4), which go into rate equation (5). The step in orange is optional (2*). It correlates the activity with the out-degree of the input signed network, i.e. the agent with the highest out-degree will generate more links. The last step (6) constitutes the final rate calculation, where we also utilize the input signed network.

reddit

Reddit is the social news aggregation website. Users can post content while others can rate and discuss it (see Fig. 1). We can distinguish communities, so-called *subreddits*, which gather people with similar opinions and beliefs. An example post is presented in Fig. 1. User submitted content on one subreddit and included hyperlink to another. We denote such a hyperlink as an interaction between two subreddits. To study such interactions among subreddits we use dataset from SNAP repository [1]. The Reddit Hyperlink Network (RHN) represents the directed connections between subreddits. A link (hyperlink) originates from a post in the source community and links to a post in the target community. The network is directed, signed and temporal.

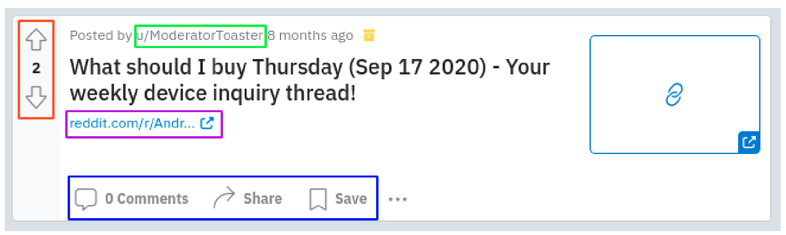

Fig. 1: Exemplary post from `r/MoronicMondayAndroid` subreddit. The post was submitted by user named `u/ModeratorToaster` (green rectangle). Currently it has two upvotes (red rectangle) and zero comments (blue rectangle). There is a hyperlink to `r/Android` subreddit marked by a violet rectangle.

Each hyperlink is endowed with a sentiment value, resulting in signed interactions. However, we are more interested in signed relations. Thus, we have to convert signed interactions into relations. The first basic approach is to set a threshold and convert averaged sentiment into plus and minus links (Fig. 2).

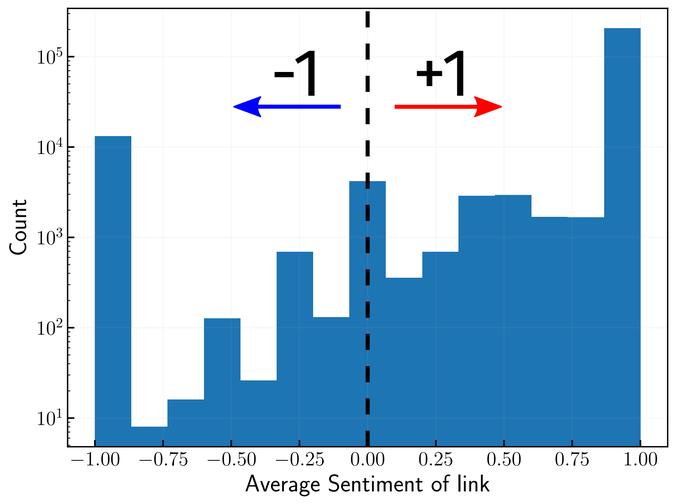

Fig. 2: Histogram of average sentiments of links. The dashed line indicates threshold used to create signed network.

Model and methods

Activity Driven Network model (ADN) was proposed to study highly dynamical networks [2]. It uses the activity potential, which describes the willingness of agents to create links. The algorithm proceeds with generating an empty graph of size N and assigning an activity potential x_i to each agent. The potential activity is usually drawn from a power-law distribution $F(x)$. At each discrete time step t with probability $a_i \Delta t$, where $a_i = \eta x_i$ each vertex i becomes active and generates m links to randomly selected nodes. Then, at the next time step $t + \Delta t$, all edges in the network are deleted and the process runs once again. We propose a few extensions: **(i)** use the input signed network to generate both the activity and preference distributions, **(ii)** use signed network as agent's preference (see Fig. 3).
We also incorporate the copying mechanism which accounts for neighbours of neighbours. It takes into account paths of length 2 between agents i and j.

$$\text{copying}_{i,j}(t) = \frac{1}{\min(k_i, k_j)} \mathbf{A}^2(t)_{i,j} \tag{1}$$

and $\text{copying}_{i,i}(t) = 0$, where $\mathbf{A}(t)$ is the accumulated undirected adjacency matrix at time t. One can define the total rate that the node j obtains a link from the active node i as follows:

$$\text{rate}_{i,j}(t) = \alpha \cdot \text{pref}_j + \beta \cdot f(\text{sign}_{i,j}) + \gamma \cdot \text{copying}_{i,j}(t-1), \tag{2}$$

where α, β and γ are constants which correspond to the strength of preferential term, sign term and copying mechanism term, respectively. All constants are in range $[0, 1]$. Function f is defined as

$$f(\text{sign}_{i,j}) = \begin{cases} 1, & \text{if the link } (i,j) \text{ is positive.} \\ 0, & \text{if the link } (i,j) \text{ is negative.} \end{cases} \tag{3}$$

Here we focus on the parts marked by green circles (see Fig. 3), i.e. whether to correlate activity and test the impact of the input signed network.

	RHN	$\beta = 0$, no corr	$\beta = 0$, corr	$\beta = 1$, no corr	$\beta = 1$, corr
Δ	1329278	747599	711171	642960	1260899
$\langle C \rangle$	0.1844	0.4878	0.4989	0.4983	0.4198
$\leftrightarrow$	29282	94950	96916	96364	71748

Tab 1. Summary statistics of the models with and without input signed network and the empirical data. Δ – number of triangles, $\langle C \rangle$ – average clustering coefficient, $\leftrightarrow$ – number of reciprocal links.

Impact of the input signed network

For the model without input signed network ($\beta = 0$) and with activity correlation (blue line) the maximum link weight matches the corresponding value from RHN (Fig. 4).

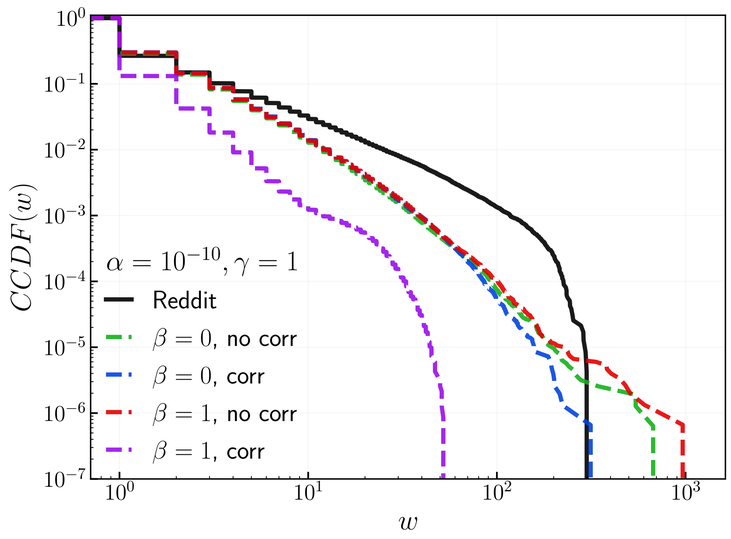

Fig. 4: Link-weight distribution for ADN models with and without input signed network.

The model with the input signed network ($\beta = 1$) and activity correlation (in violet) in 6 out of 7 cases has the highest number of motifs that are closer to empirical data than the rest of the models (Fig. 5).

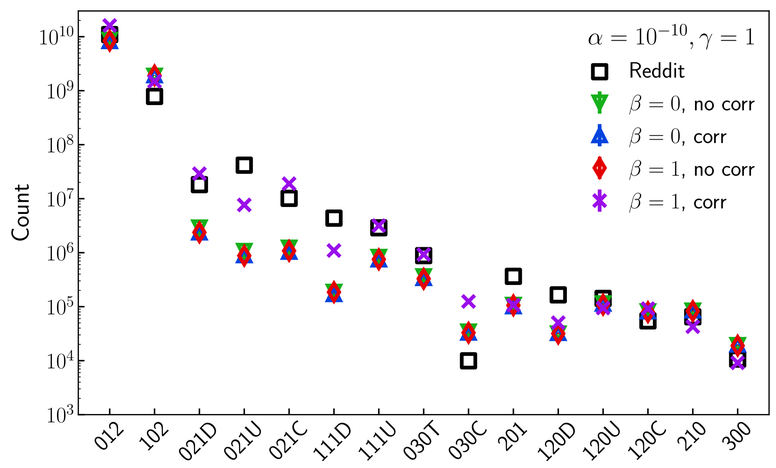

Fig. 5: Triad census for models with and without input signed network.

Conclusions

- We reveal the possibility of developing the ABM to generate interactions from agents' relations, their preference and their memory about the neighbours of neighbours.
- The results lead to the conclusion of only a slight impact of the input signed network on the $P(w)$ distribution.
- We successfully restore a similar number of motifs' types by analyzing the triad census.

References

[1] Kumar, Srijan et al. "Community interaction and conflict on the web." Proceedings of the 2018 world wide web conference.

[2] Perra, N. et al. "Activity driven modeling of time varying networks". Scientific Reports 2. 1(2012): 1–7.

Acknowledgments

The work was developed withing the scope of the ALPHORN project "Signed Relations and Structural Balance in Complex Systems: From Data to Models" funded from National Science Centre, Poland Grant No. 2019/01/Y/ST2/00058

Contact information

Robert Jankowski
Faculty of Physics, Warsaw University of Technology, Poland
E-mail: **robert.jankowski3.stud@pw.edu.pl**