Statistical cyber-security

Modelling new edge formation in large computer networks

Francesco Sanna Passino
francesco.sanna-passino16@imperial.ac.uk

Dr Nicholas Heard
n.heard@imperial.ac.uk

Imperial College London, Department of Mathematics

1. Problem

Monitoring and detecting anomalies in computer networks is an extremely challenging task. The quantity of data available is massive and large networks are constantly target of attacks from potential intruders. It is possible to employ advanced techniques, based on statistical models, in order to identify suspicious patterns within the network. In particular, in this project we focus on modelling the network graph with two main purposes in mind:

- predicting future links
- identifying anomalous edges

We use the NetFlow data collected by Imperial College London. Each record represents a connection and contains information such as: source IP, destination IP, source and destination ports, bytes and packets sent, duration of the connection event.

2. Graph representation

Given a set of Netflow records within a given time interval, we can construct a directed graph $\mathbb{G} = (V_c, V_s, E)$ where:

- V_c is the set of clients, $|V_c| = n_c$,
- V_s is the set of servers, $|V_s| = n_s$,
- E is the edge set, containing dyads $(i, j),\ i \in V_c,\ j \in V_s$.

We draw an edge if a client $i \in V_c$ connects to server $j \in V_s$ within the time interval, and we write $(i, j) \in E$.

From $\mathbb{G}$, we can obtain a rectangular adjacency matrix $\mathbf{A} = \{A_{ij}\}$, of dimension $n_c \times n_s$. We have:

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{array}{cc} & \text{Servers} \\ \text{Clients} & \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 1 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & 1 & \cdots & 1 & 1 \end{pmatrix} \end{array}$$

Note that this object is hugely sparse.
It is also useful to consider a weighted version $\mathbf{W} = \{W_{ij}\}$ of the rectangular adjacency matrix. The weights associated with each dyad $(i, j) \in E$ are obtained as follows:

$$W_{ij} = \log(1 + N_{ij})$$

where N_{ij} is the number of connections between two nodes within the time interval.

3. Exploratory Data Analysis

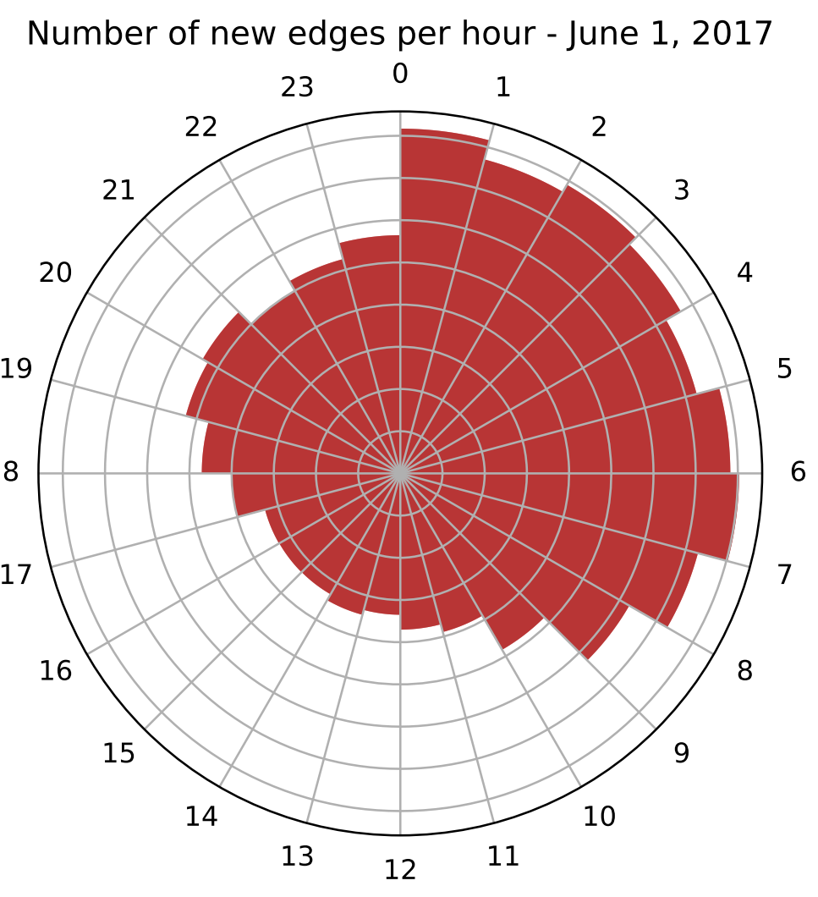

Figure 1: *New edges per hour on June 1, 2017. First bar:* $16{,}336{,}104$ *edges (all new). Lowest bar (2-3pm):* $6{,}614{,}870$ *new edges.*

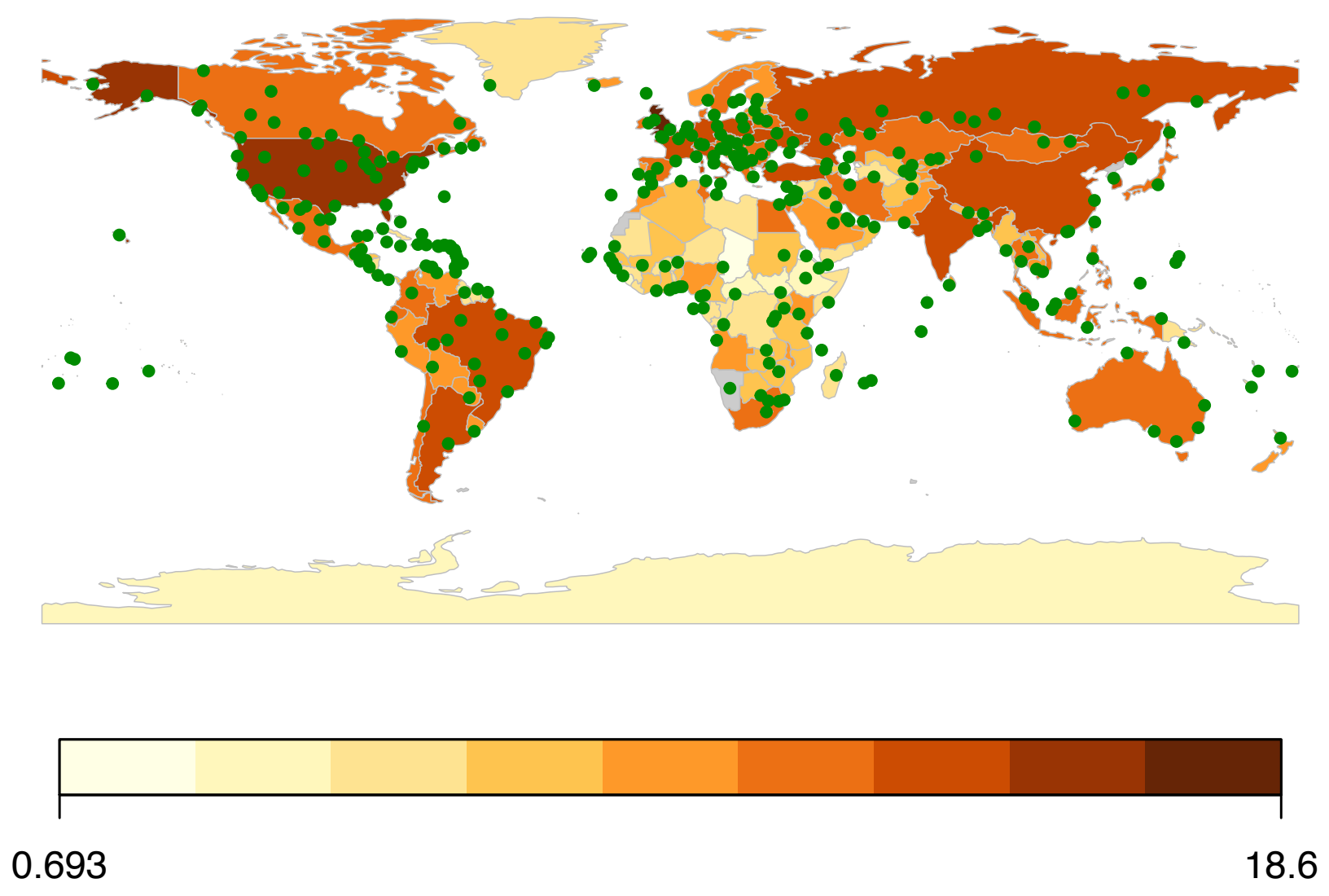

Figure 2: *Client locations on June 1, 2017, 10-11am, intensity given by* $\log(1 + N_i)$*, where* N_i *is the number of connections of* c_i*.*

4. Some features of the Imperial College network

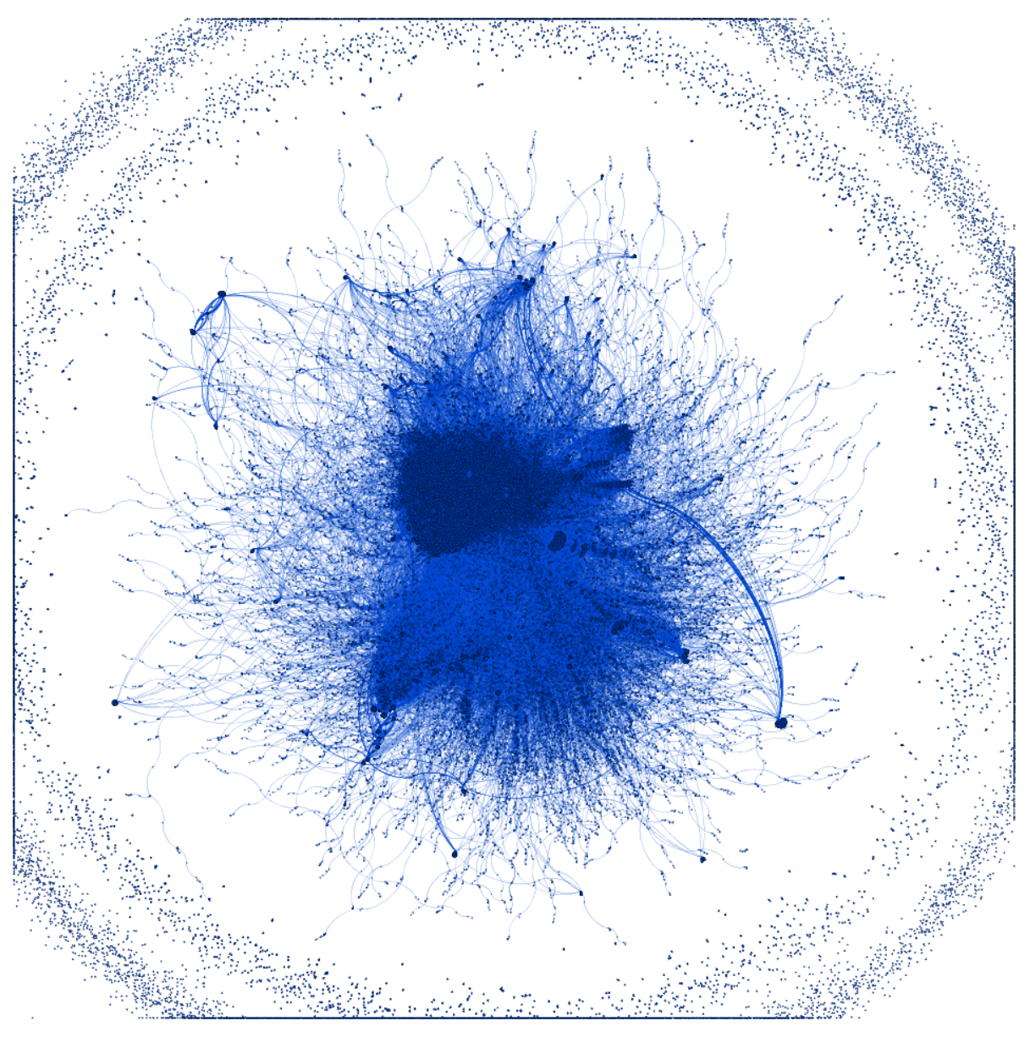

Figure 3: *Network on June 7, 2017, 11:15-11:16am.*

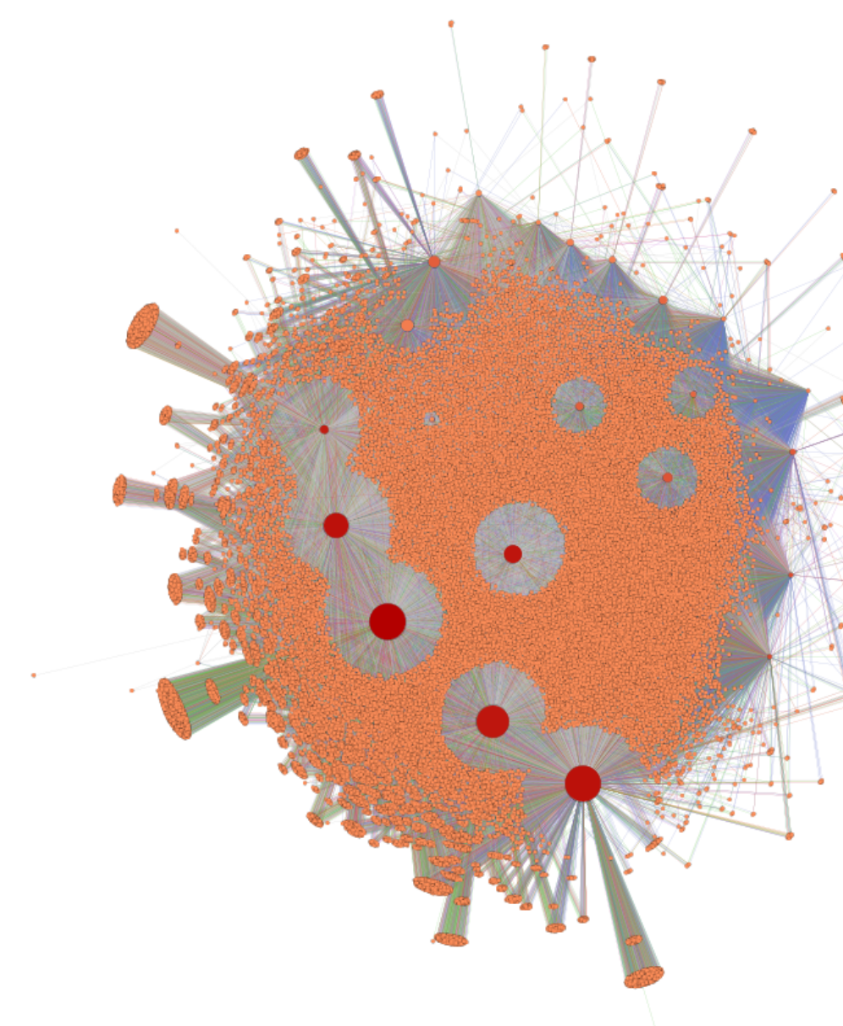

Figure 4: *Network on June 1, 2017, 10-11am, IPs grouped by two leftmost octets.*

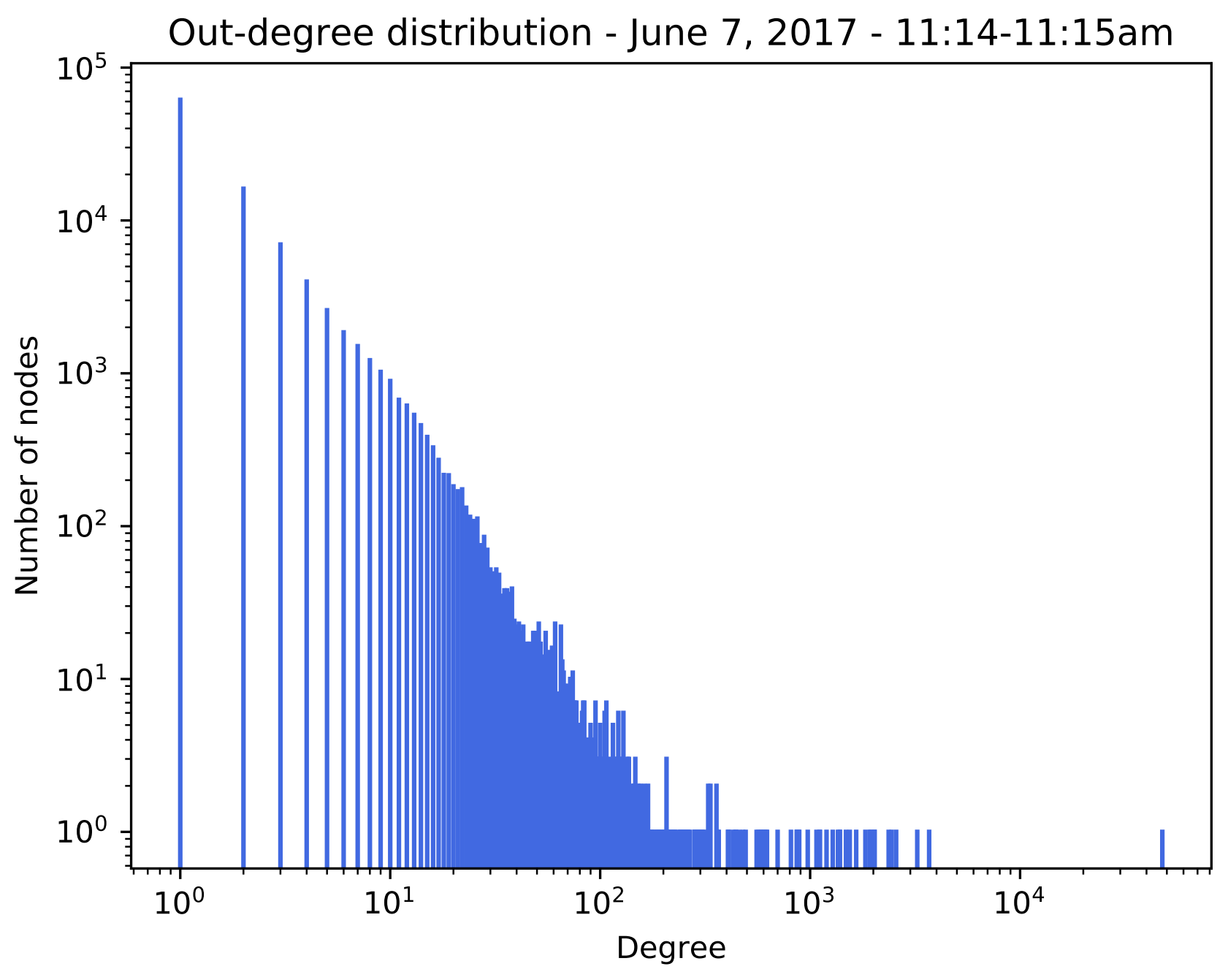

Figure 5: *Out-degree distribution on June 7, 2017, 11:14-11:15am.*

5. Methods

Aim: starting from a binary adjacency matrix $\mathbf{A}$ or from its weighted version $\mathbf{W}$, construct a matrix of scores $\mathbf{S} = \{S_{ij}\}$, and use the ranked values of the S_{ij}'s in a binary classification test, in order to predict the future status of the connection between c_i and s_j. In this work, we are particularly interested in the prediction of new links, for anomaly detection purposes.

The following methods can be used to construct $\mathbf{S} = \{S_{ij}\}$:

- Truncated SVD (tSVD) (Dunlavy, Kolda and Acar, 2011):

$$\mathbf{A} \approx \mathbf{A}_r = \mathbf{U}_r \mathbf{D}_r \mathbf{V}_r^\top \implies \mathbf{S} = \mathbf{U}_r \mathbf{D}_r \mathbf{V}_r^\top$$

  where $\mathbf{U}_r$ and $\mathbf{V}_r^\top$ contain left and right singular vectors corresponding to the top r singular values, stored in the diagonal matrix $\mathbf{D}_r = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$.
- Truncated Katz scores (tKatz) (Dunlavy, Kolda and Acar, 2011):

$$\mathbf{S} = \mathbf{U}_r \mathbf{\Psi}_r^- \mathbf{V}_r^\top$$

  where $\mathbf{U}_r$ and $\mathbf{V}_r^\top$ are the same matrices as above, and $\mathbf{\Psi}_r^- = \mathrm{diag}(\psi_1^-, \dots, \psi_r^-)$ with $\psi_i^- = \beta\sigma_i/(1 - \beta^2\sigma_i^2)$.
- Truncated Eigen-Decomposition (TED) (Rubin-Delanchy, Adams and Heard, 2016): let $\tilde{\mathbf{A}}_{\text{sym}} = \{\tilde{A}_{ij}^{\text{sym}}\}$ represent the squared symmetric adjacency matrix for the undirected graph. Its normalised modified Laplacian is:

$$\tilde{\mathbf{L}}_{\text{sym}} = \mathbf{D}^{-\frac{1}{2}} \tilde{\mathbf{A}}_{\text{sym}} \mathbf{D}^{-\frac{1}{2}}$$

  where $\mathbf{D}$ is the degree-matrix $\mathbf{D} = \mathrm{diag}(d_1, \dots, d_{|V|})$, $d_i = \sum_{j=1}^{|V|} \tilde{A}_{ij}^{\text{sym}}$. The spectral decomposition gives $\tilde{\mathbf{L}}_{\text{sym}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, which can be truncated using the r largest eigenvalues, or the top r eigenvalues in magnitude:

$$\mathbf{S}_r = \mathbf{Q}_r \mathbf{\Lambda}_r \mathbf{Q}_r^\top \qquad \text{or} \qquad \mathbf{S}_{|r|} = \mathbf{Q}_{|r|} \mathbf{\Lambda}_{|r|} \mathbf{Q}_{|r|}^\top$$

- A popular link probability model used in the network literature (for example, Caron and Fox, 2014) is:

$$\mathbb{P}(A_{ij} = 1 | w_i, w_j') = 1 - \exp\left\{-w_i w_j'\right\} \qquad (1)$$

  where $\{w_i\}_{i=1}^{n_c}$ and $\{w_j'\}_{j=1}^{n_s}$ are sociability parameters for clients and servers respectively. The sociabilities can be approximated using functions of in-degree d_j^{in}, and out-degree d_i^{out}:

$$\hat{\mathbb{P}}(A_{ij} = 1) = 1 - \exp\{-\log(d_i^{\text{out}} + 1)\log(d_j^{\text{in}} + 1)\}$$

- Novel score for bipartite graphs, based on the idea of *"neighbouring"*: overlap statistic. For k levels of nesting:

$$\mathbf{S}^{(k)} = \begin{cases} \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{\frac{k-1}{2}}\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{\frac{k-1}{2}}\mathbf{A} & k \in \mathbb{N}_{\text{odd}} \\ (\mathbf{A}\mathbf{A}^\top)^{\frac{k}{2}}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{\frac{k}{2}} & k \in \mathbb{N}_{\text{even}} \end{cases}$$

- Sparse graph using exchangeable random measures model (Caron and Fox, 2014): the results of the MCMC sampler for the sociability parameters $\{w_i\}_{i=1}^{n_c}$ and $\{w_j'\}_{j=1}^{n_s}$ in the bipartite GGP model can be used to form scores $\hat{\mathbb{P}}(A_{ij} = 1)$ using (1).

6. Results

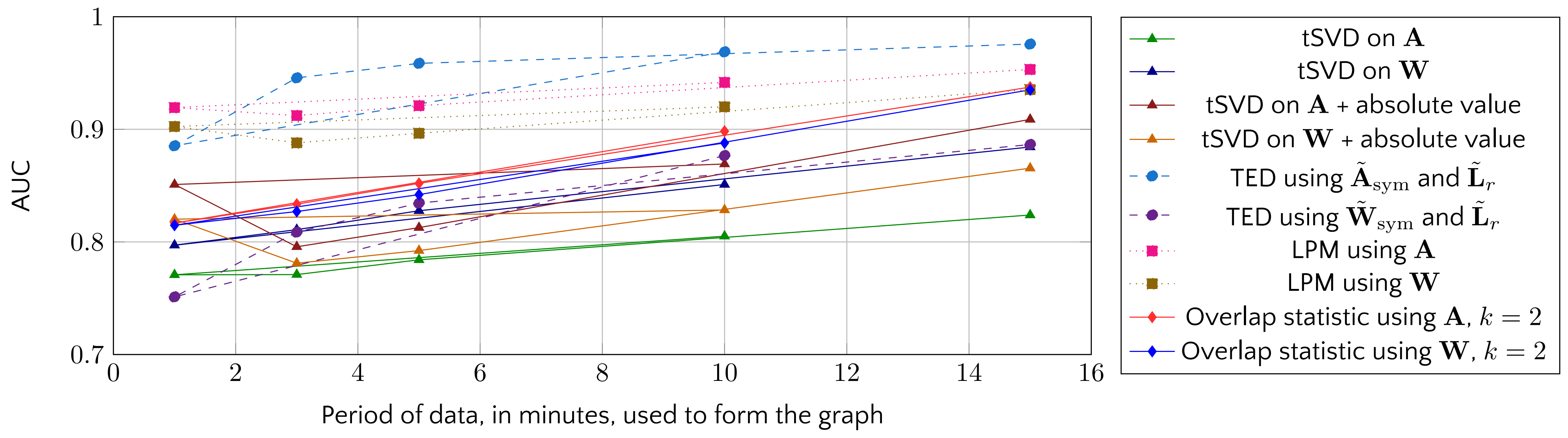

Figure 6: *AUC scores obtained from the link prediction procedure of new links on the graph obtained on June 7, 2017, 11:15-11:16am, using different predictor graphs, constructed from data for 1, 3, 5, 10 and 15 minutes before the time of interest,* $r = 30$, $\beta = 0.001$.

References

■ Caron, F. and E.B. Fox (2014). "Sparse graphs using exchangeable random measures". In: ArXiv e-prints. arXiv: 1401.1137.

■ Dunlavy, D.M., T.G. Kolda, and E. Acar (2011). "Temporal link prediction using matrix and tensor factorizations". In: ACM Transactions on Knowledge Discovery from Data (TKDD) 5(2).

■ Rubin-Delanchy, P., N.M. Adams, and N.A. Heard (2016). "Disassortativity of computer networks". In: 2016 IEEE Conference on Intelligence and Security Informatics (ISI), pp. 243–247.

7. Conclusions

■ The performances are remarkable given that only a small subset of edges is common between the graphs.
■ Considering larger graphs, and so more information, clearly improve the prediction performance.
■ No clear improvement when the weighted version of the adjacency matrix is used.
■ Best performance consistently achieved by the TED of the normalised Laplacian matrix, using the largest graph.
■ For the prediction based on one minute of data only, the model of Caron and Fox (2014) and the approximation of the link probability in (1) using in-degrees and out-degrees have the best performance.