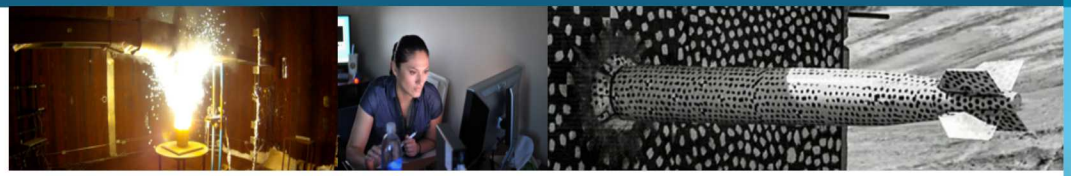
SAND2020-xxxx

SAND2020-4720PE

Sandia National Laboratories

Partitioning Communication Streams into Graph Snapshots



Rich Field

Data Science & Applications

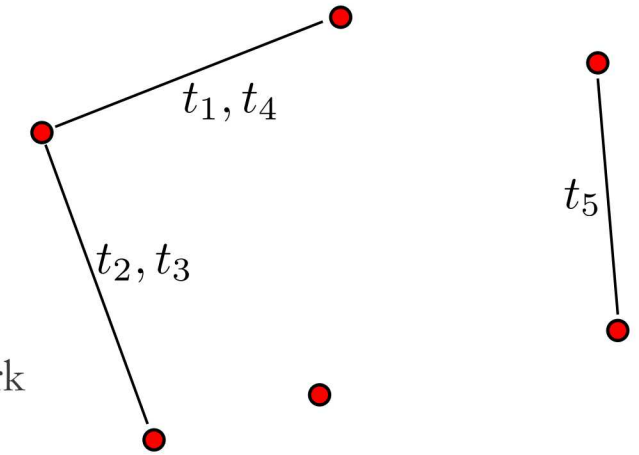
rvfield@sandia.gov

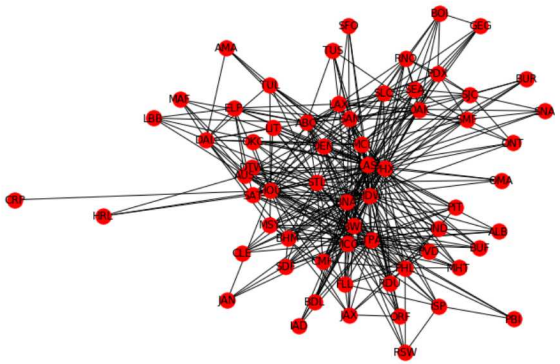
Co-authors: J. Wendt and C. Phillips (SNL); T. Wilson (Cornell); S. Soundarajan (Syracuse Univ.); S. Bhowmick (Univ. North Texas)

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

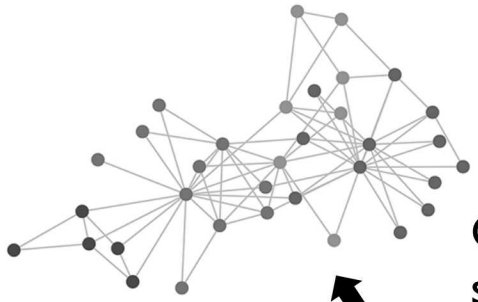
Motivation

- Graphs – models used to represent systems of interacting entities
  - Nodes represent entities
  - Edges represent interactions
- These systems are often dynamic
  - Interactions change over time
  - Examples: Emails, text messages, computer network traffic
  - Often we build static graphs
- A key question – how can we appropriately build a static graph from dynamic data?
  - We will use a subcollection and create a graph "snapshot"
  - If the subcollection is too small the graph is incomplete
  - If the subcollection is too large, the graph may obscure some important interactions

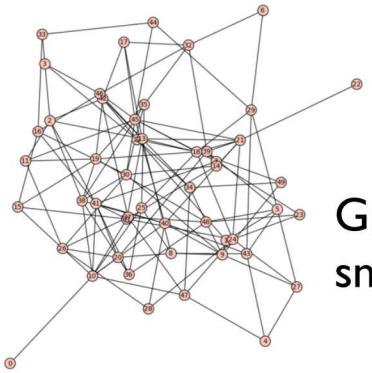




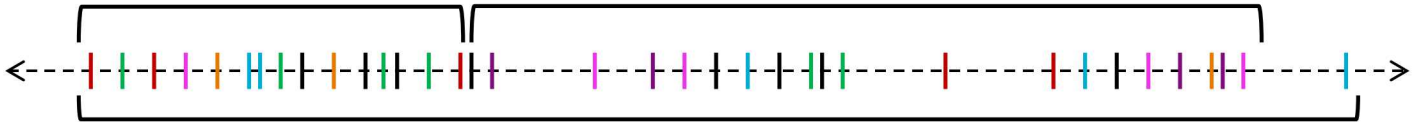
Balance Over and Undersampling



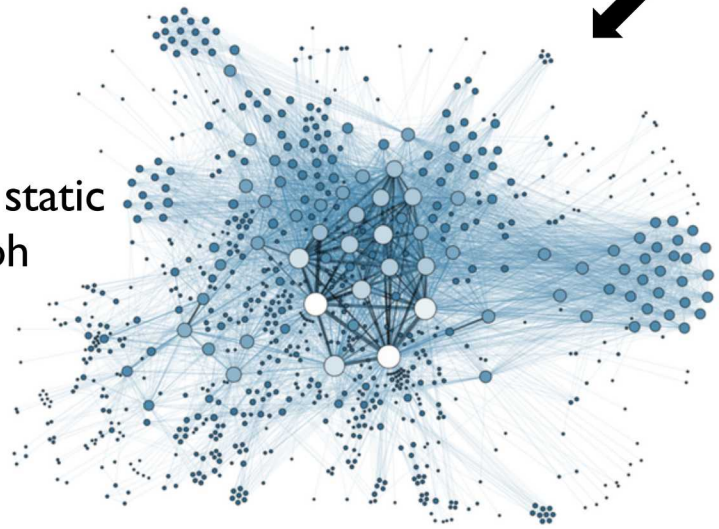
Graph snapshot 1



Graph snapshot 2



Communication events over time



Full static graph

- The nature of a network can change due to external events
  - Nature, economics, health
  - Entities join and leave (and may join again)
  - Possible to infer relationships that may not exist
- We want snapshots that are:
  - Large enough to reasonably represent graph activity (no undersampling)
  - Small enough to stop at major network behavior changes (no oversampling)

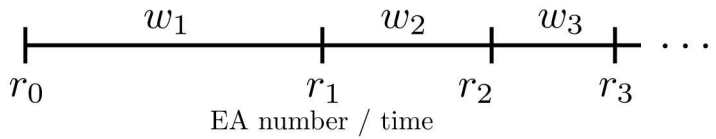
Outline

- We need a rigorous way of defining appropriate subcollections
  - Edge Advertisements into Snapshots using Evolving Expectations (EASEE)
- Datasets
- Model
  - Edge advertisement (EA) types
  - Modeling the probability of EA types
  - Near-term graph growth
  - Partitioning streams into snapshots
- Synthetic EA data
- Results
  - Synthetic data
  - Real datasets
  - Compared EASEE with existing algorithm (not discussed)
- Related work

Edge Advertisements into Snapshots Using Evolving Expectations (EASEE)

- Edge advertisement (u, v, t)
  - Communication takes place between entities u and v at time t
  - We assume a streaming collection of EAs
  - An edge may advertise more than once
- Graph snapshots
  - A graph built from all EAs within a specified interval of size w_j
  - Extend to a sequence of graph snapshots $\{(\mathcal{V}_j, \mathcal{E}_j), j = 1, 2, \ldots\}$

$$\mathcal{E}_j = \{(u, v) \colon (u, v, t)_i, i = r_{j-1} + 1, \ldots, r_j\}$$
$$\mathcal{V}_j = \{u \colon (u, v) \in \mathcal{E}_j \text{ or } (v, u) \in \mathcal{E}_j\}$$

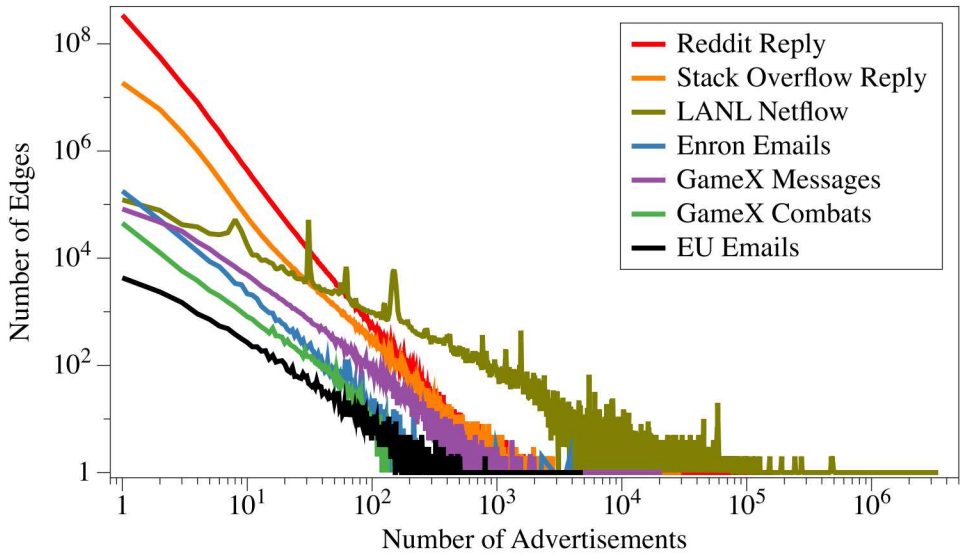


- The EASEE algorithm:
  - Predicts (near) future graph size
  - Uses prediction to find a smallest sufficient snapshot (not undersampled)
  - Possibly combines two or more sufficiently similar neighboring snapshots
  - Executes in real time

Datasets

- *EU Core Emails*: A common, often repeating, type of communication
- *Enron Emails*: Emails for approx. 150 users at Enron Corp.
- *GameX*: Logs of *Combats* and *Messages* for an online game
- *Stack Overflow Reply*: Log of questions and answers
- *Wiki-Vote*: Administrator elections and vote history
- *Reddit Reply*: Log of posts and comments on social news
- *LANL Netflow*: 32 days of computer network traffic (both human and automated)
- Also consider synthetic data

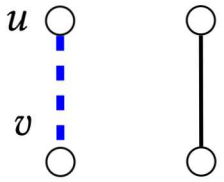
Name	Number of EAs	Number of Nodes	Number of Edges
EU Core Emails	327,336	986	16,064
Enron Emails	1,283,755	84,511	316,061
GameX Combats	500,327	10,589	86,351
GameX Messages	4,515,396	22,442	293,860
Stack Overflow Reply	63,496,479	2,601,977	29,541,284
Reddit Reply	646,024,723	8,901,033	437,747,667
LANL Netflow	2,585,934,400	166,925	1,237,992



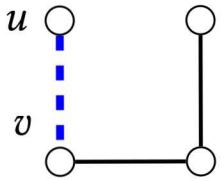
Edge Advertisement "Types"

- Suppose an edge (u, v) advertises at time t. It can be one of only four types.
- *Type N2*: New edge with 2 new nodes
  - Neither u nor v were part of any edge advertisement prior to time t
- *Type N1*: New edge with 1 new node
  - Exactly one of u or v were part of at least one edge advertisement prior to time t
- *Type N0*: New edge with 0 new nodes
  - Both u and v were part of at least one edge advertisement prior to time t
- *Type R*: Repeat edge
  - Edge (u, v) advertised at least once prior to time t
  - This is a special case of type N0
- Note: we can consider these types for each snapshot

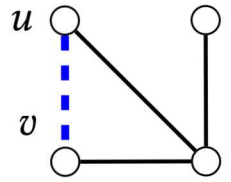
N2-type



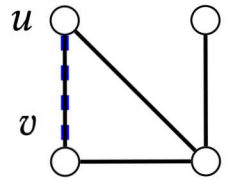
N1-type



N0-type



R-type



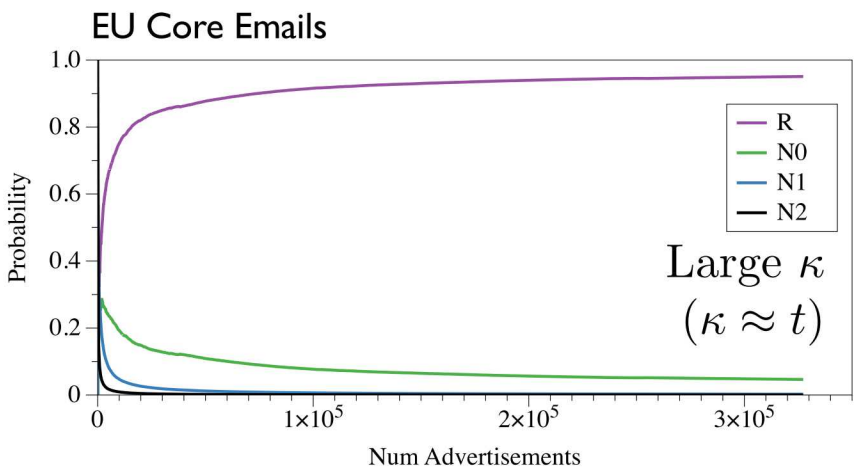
Edge Advertisement Type Probabilities

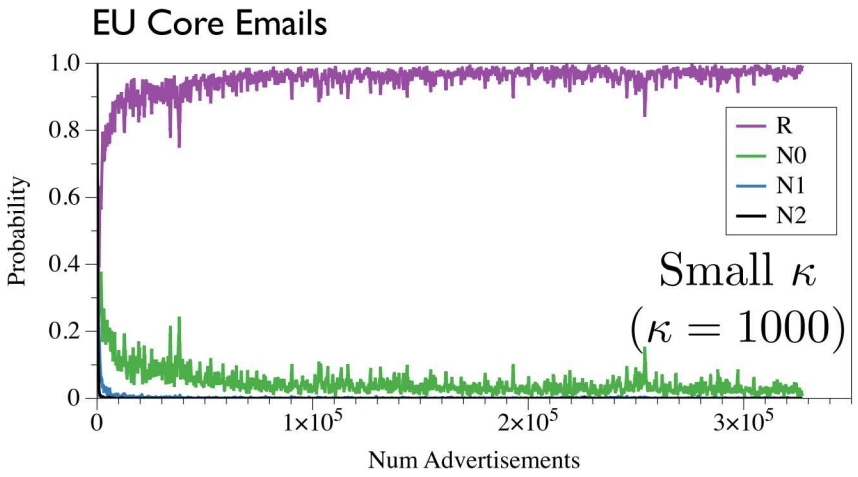
- For edge advertisement (u, v, t) determine the probabilities

$$p_i(t) = \Pr((u, v, t) \text{ is type } i), \ i \in \{\text{N2}, \text{N1}, \text{N0}, \text{R}\}$$

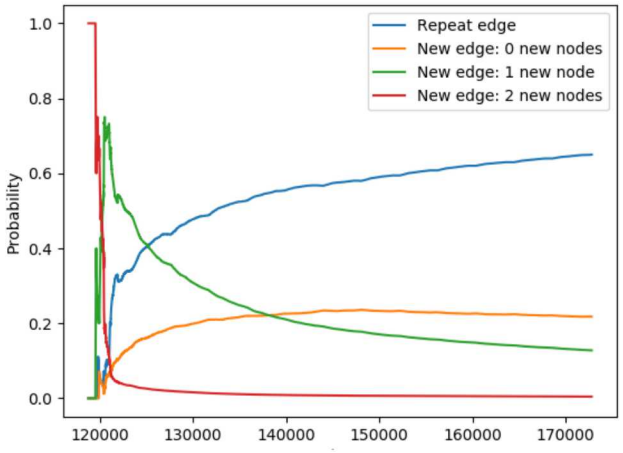
- Two approaches
  - Counting over a sliding window

$$p_i(t; \kappa) \approx \frac{1}{\kappa} \sum_{\tau=t-\kappa}^{t} \mathbf{1}\left((u, v, \tau) \text{ is type } i\right)$$





Parameterized Model For EA Type Probabilities (1 of 2)



- We observe
  - Probabilities approach steady-state values
  - Monotonic behavior in R and N2
    - Type R becomes more likely over time
    - Type N2 becomes less likely over time
- Model

$$
\begin{aligned}
p_{\mathrm{R}}(t) &= \alpha\left(1-e^{-a\,t}\right)\\
p_{\mathrm{N0}}(t) &= \gamma\left(1-e^{-c\,t}\right)\left(1-\alpha\left(1-e^{-a\,t}\right)\right)\\
p_{\mathrm{N1}}(t) &= \left(1+\gamma\left(e^{-c\,t}-1\right)+\beta\left(e^{-b\,t}-1\right)-e^{-b\,t}\right)\times\\
&\qquad \left(1-\alpha\left(1-e^{-a\,t}\right)\right)\\
p_{\mathrm{N2}}(t) &= \left((1-\beta)e^{-b\,t}+\beta\right)\left(1-\alpha\left(1-e^{-a\,t}\right)\right)
\end{aligned}
$$

- Parameters (fit to data using least squares)

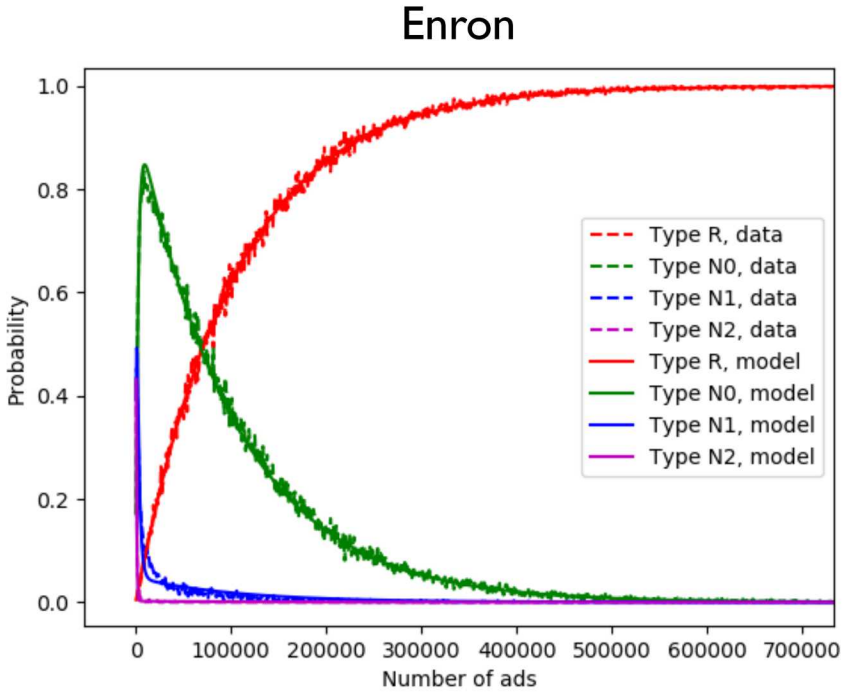
α, β, γ : steady state values

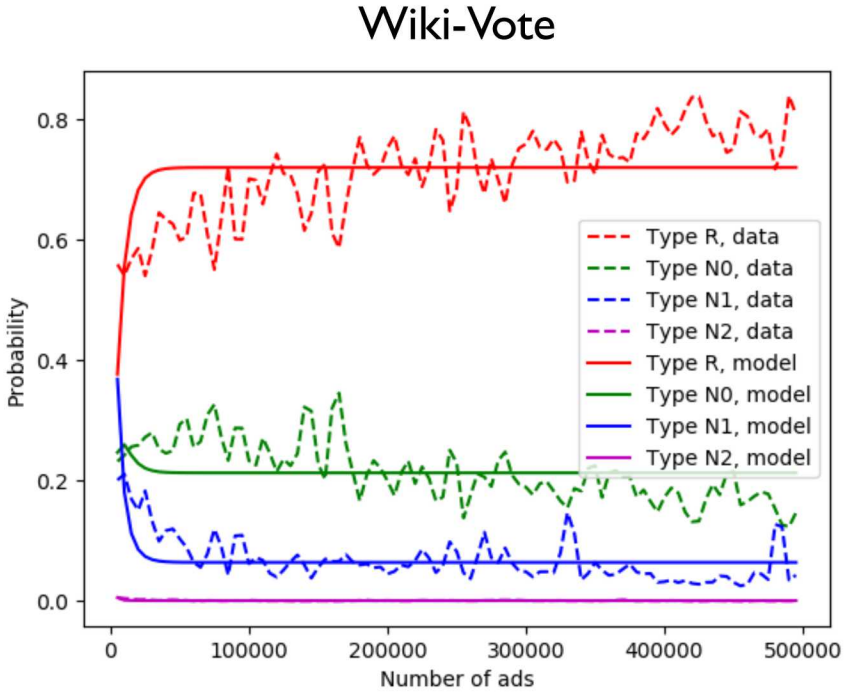
a : rate of repeat edges

b : rate that two new nodes join the network

c : rate of new communications for established members

Parameterized Model For EA Type Probabilities (2 of 2)





Forecasting Graph Evolution – Number of Edges

- Let m_k denote the number of edges in the graph after k edge advertisements (EAs)

$$m_{k+1} = \begin{cases} m_k & \text{with probability } p_{\mathrm{R}}(k) \\ m_k + 1 & \text{with probability } (1 - p_{\mathrm{R}}(k)) \end{cases}$$

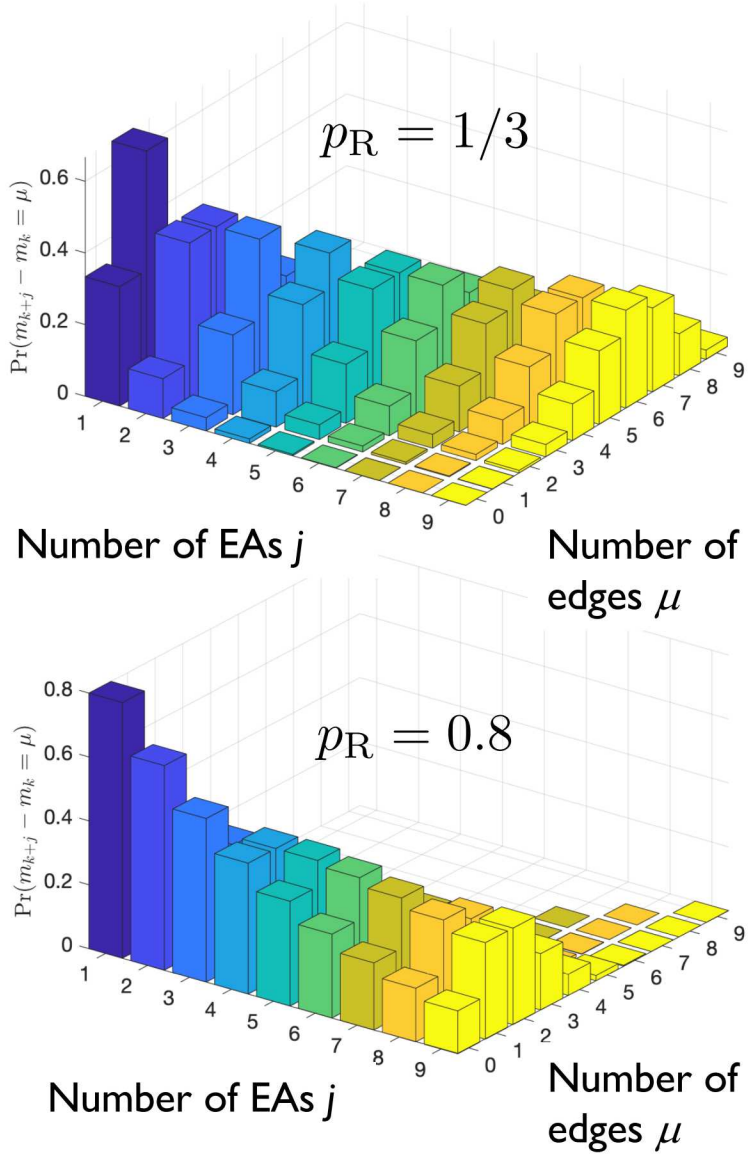
- Forecast number of edges $j > 0$ EAs ahead
  - Assume the probability of EA type R has reached a steady-state value, that is

$$p_{\mathrm{R}}(k) = p_{\mathrm{R}}(k+1) = \cdots = p_{\mathrm{R}}(k+j) = p_{\mathrm{R}}$$

  - Probability of exactly μ additional edges at $j > 0$ EAs ahead follows a binomial distribution

$$\Pr(m_{k+j} - m_k = \mu) = \binom{j}{\mu} p_{\mathrm{R}}^{j-\mu}(1 - p_{\mathrm{R}})^{\mu}$$

Mean is $j(1 - p_{\mathrm{R}})$



Forecasting Graph Evolution – Number of Nodes

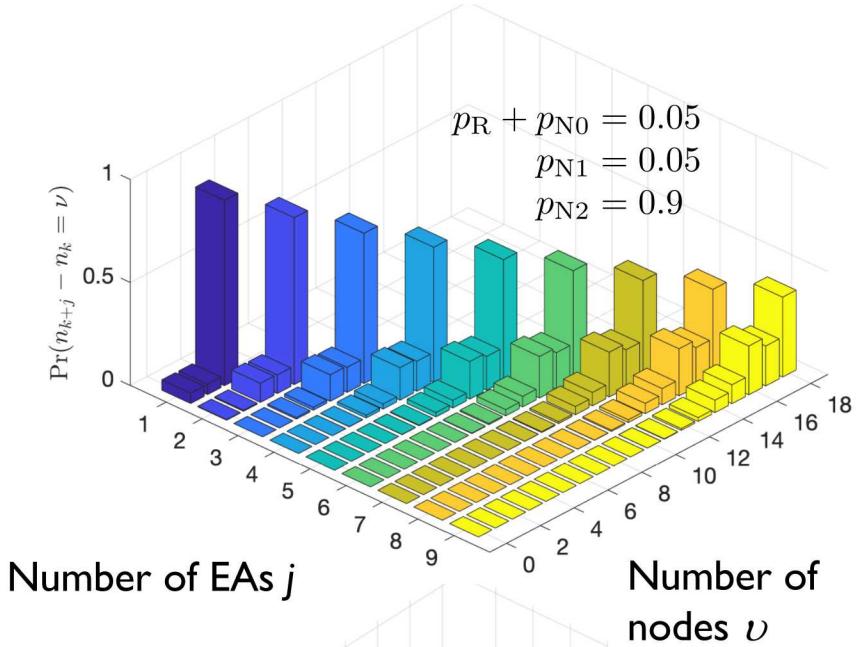
- Let n_k denote the number of nodes in the graph after k edge advertisements (EAs)

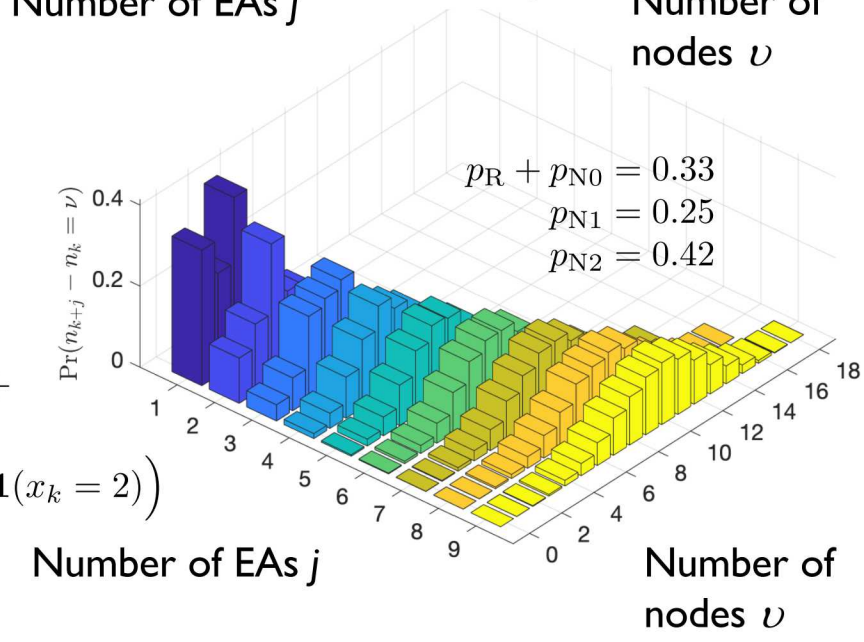
$$n_{k+1} = \begin{cases} n_k & \text{with probability } p_{\mathrm{R}}(k) + p_{\mathrm{N0}}(k) \\ n_k + 1 & \text{with probability } p_{\mathrm{N1}}(k) \\ n_k + 2 & \text{with probability } p_{\mathrm{N2}}(k) \end{cases}$$

- Forecast number of nodes $j > 0$ EAs ahead
  - Assume all edge type probabilities have reached a steady-state value
  - Probability of exactly υ additional edges at $j > 0$ EAs ahead

$$\Pr(n_{k+j} - n_k = \nu) = \sum_{\substack{x_1,\ldots,x_j \in \{0,1,2\} \\ x_1 + \cdots + x_j = \nu}} \prod_{k=1}^{j} \Big((p_{\mathrm{R}} + p_{\mathrm{N}})\mathbf{1}(x_k = 0) + p_{\mathrm{N1}}\mathbf{1}(x_k = 1) + p_{\mathrm{N2}}\mathbf{1}(x_k = 2) \Big)$$

Mean is $j(p_{\mathrm{N1}} + 2\,p_{\mathrm{N2}})$





Partitioning the Temporal Stream of EAs

- Identify the minimum sufficient size of a snapshot
  - Goal is to avoid undersampling
  - Utilize the forecast models (average number of expected nodes / edges)
  - Want "steady" graph growth
  - A snapshot interval is deemed sufficient when these expected new nodes and edges are approximately constant
    - Smoothing
    - Numerical differentiation
  - This produces the sequence of snapshots $\{(\mathcal{V}_j, \mathcal{E}_j), j = 1, 2, \ldots\}$
- Merge adjacent snapshots
  - Goal is to avoid oversampling
  - Similarity calculations on two adjacent snapshots

$$\frac{c_j \cdot c_{j+1}}{\|c_j\|_2 \, \|c_{j+1}\|_2}$$

c_j : Vector of edge (or node) counts for snapshot j

  - If this similarity metric is above a threshold, we merge snapshots j and $j + 1$

Synthetic Edge Advertisement Data

- Let G be a static graph with n nodes and m edges
  - Can be real or synthetic, but organized by community*
  - Let $A = \{a_{ij}\}$ be the n x n adjacency matrix
- Define two submatrices of A that define subgraphs

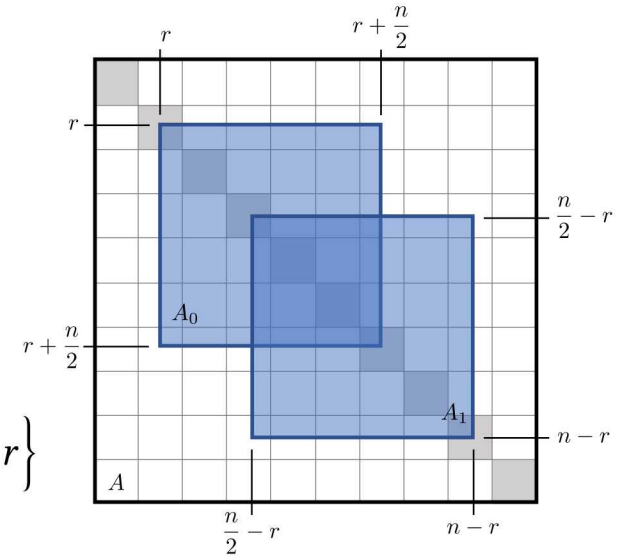
$$A_0 = \left\{a_{ij} : r \le i, j \le r + \frac{n}{2}\right\} \qquad A_1 = \left\{a_{ij} : \frac{n}{2} - r \le i, j \le n - r\right\}$$

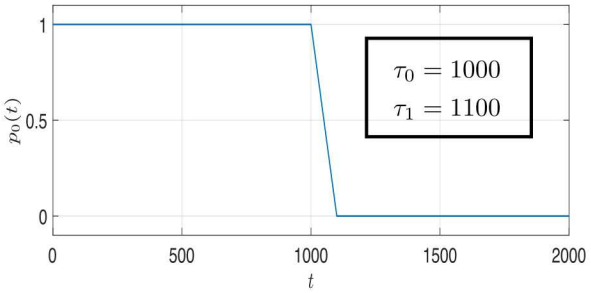
$$\text{overlap} = \frac{(2r)^2}{\left(\frac{n}{2}\right)^2} = \frac{16r^2}{n^2} \qquad r \in \{0, 1, \ldots, n/4\} \text{ is a parameter}$$

- Let $\mathcal{E}_0$ and $\mathcal{E}_1$ be edge sets defined by A_0 and A_1
- Create a sequence of edge advertisements (u, v, t)
  - Ad times t are drawn from a Poisson process with rate λ
  - Edges (u, v) are drawn from $\mathcal{E}_0$ with probability $p_0(t)$

$$\alpha = 100\frac{\lambda(\tau_1 - \tau_0)}{m} \text{ measures the rate of transition}$$

- Create data sets with various values for overlap, transition rate α, and starting time τ_0

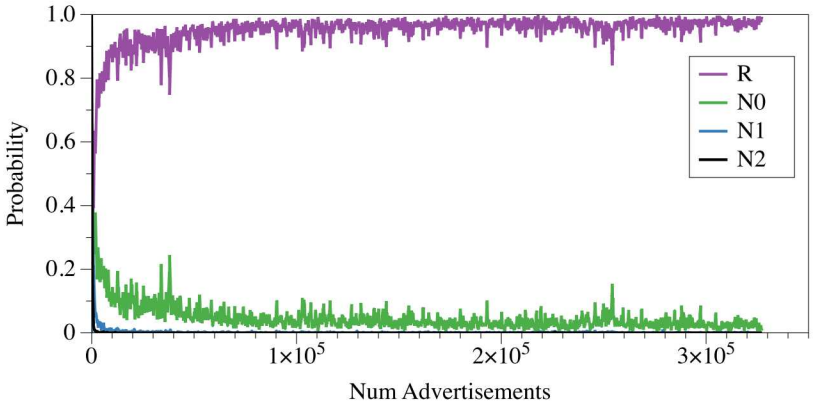


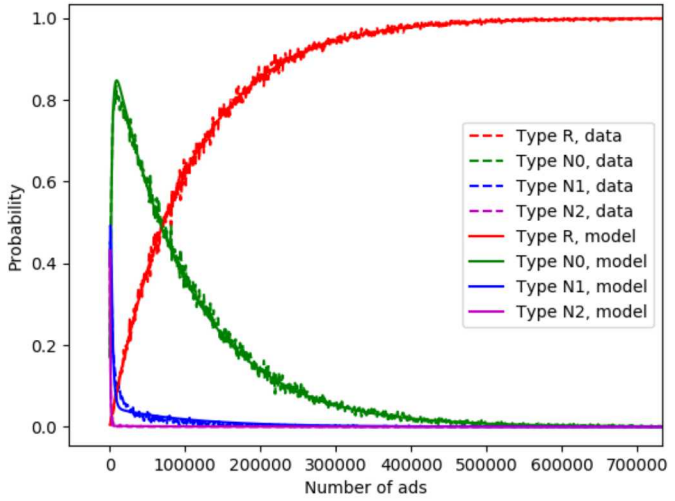


* We considered a SBM with 10 communities inter- and intra-edge probabilities equal to 0.01 and 0.05, respectively.

Results – Synthetic EA Streams (1 of 2)

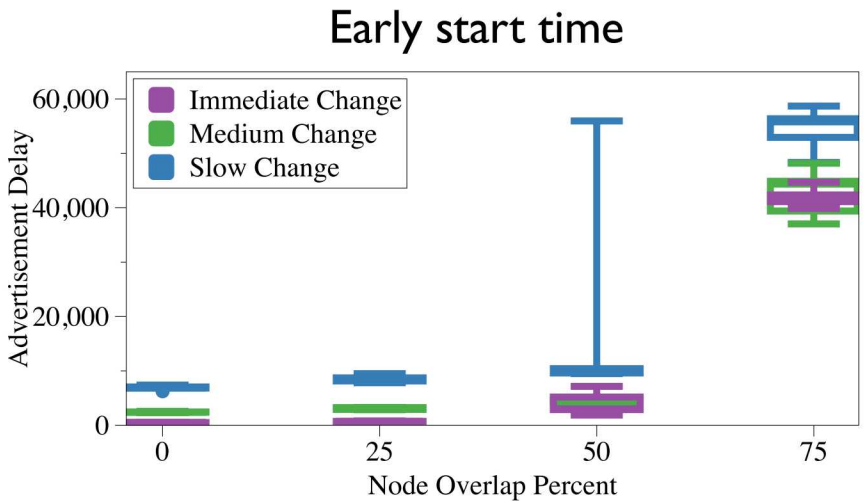
- Looking for "stable growth"
  - Edge type probabilities at near steady state
- Node overlap
  - 0, 25%, 50%, and 75%
  - Expect performance to degrade with increasing overlap
- Transition rate (slope of the curve)
  - Immediate, Medium, and Slow
  - Expect performance to degrade as transition rate slows
- Transition starting time (parameter τ_0)
  - Early, Medium, and Late
  - Changes with early starting times may be undetectable
- Considered 5 random samples for each case

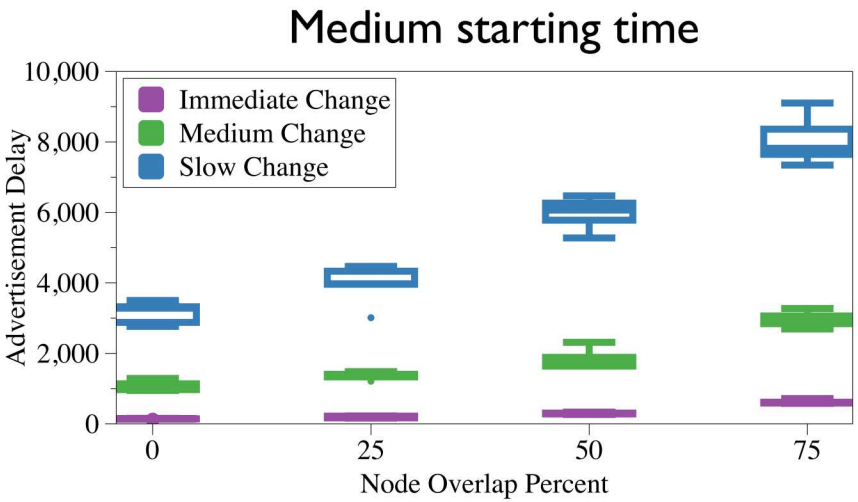


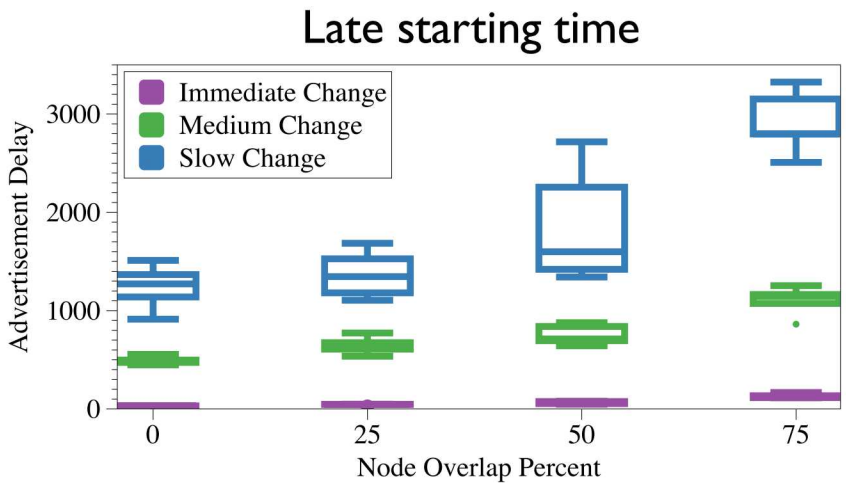


Results – Synthetic EA Streams (2 of 2)

- Horizontal axis is the overlap
- Colors are the transition rate
- Vertical axis is the error
- 3 plots for 3 different start times
  - "Early" case: the system hasn't stabilized enough to detect the change
- EASEE consistently finds one change point

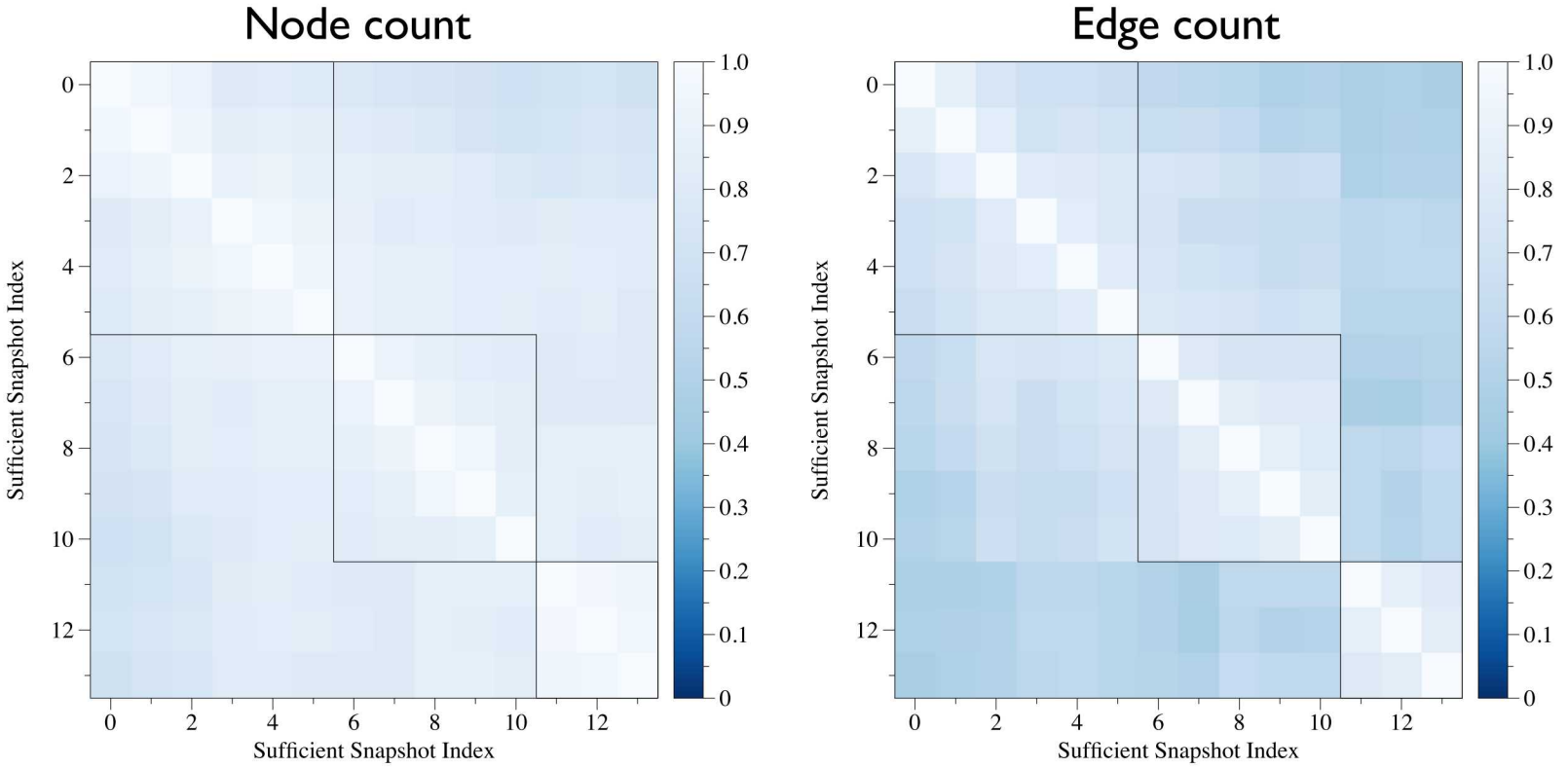






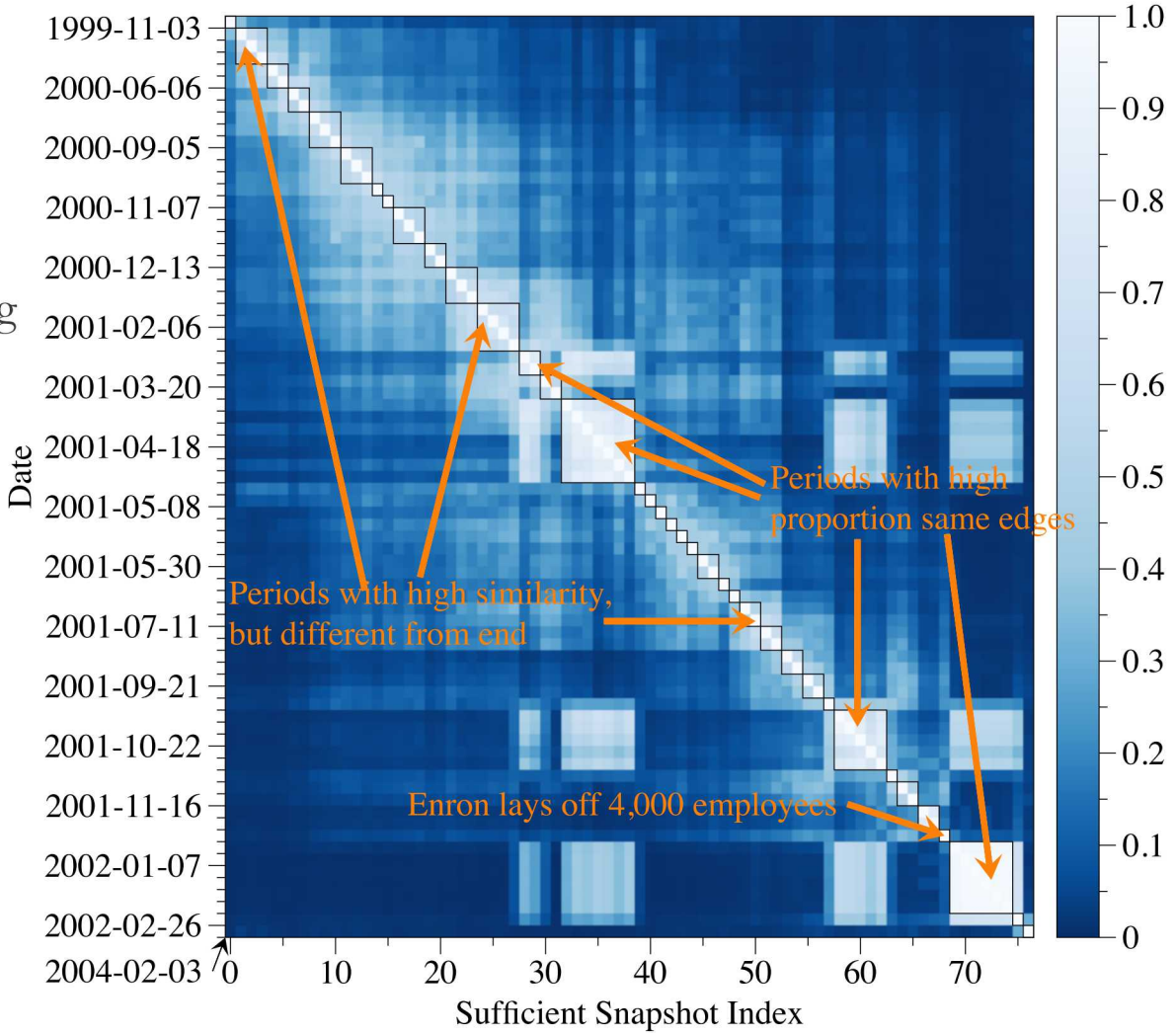
Results – EU Core Emails

- 14 sufficient snapshots produced
- Merged as (0 – 5, 6 – 10, 11 – 13)
- Similarities based on edge counts is more illustrative



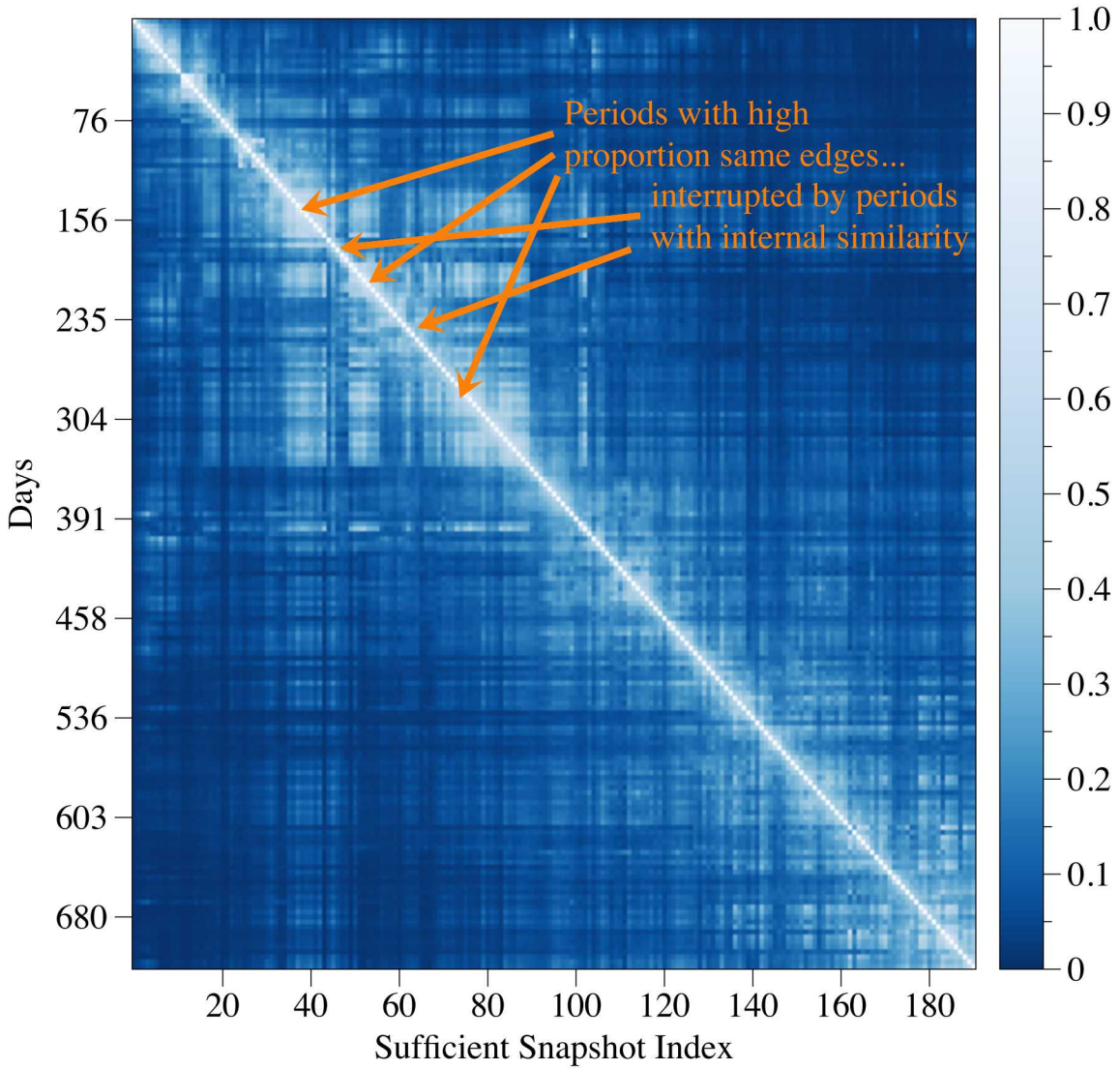
Results – Enron Emails

- Considerable changes occur over time
  - Periods of varying length with high similarity
  - There is some notable long range similarity
- 77 graph snapshots merged in 34 static graphs



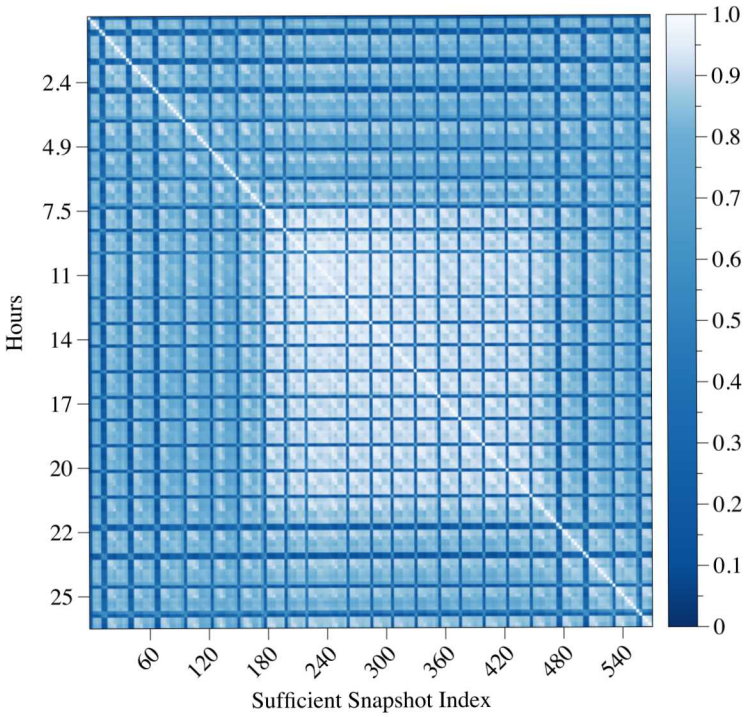
Results – GameX Messaging

- Similar results
- Approximately 200 snapshots over 2 years of messaging data

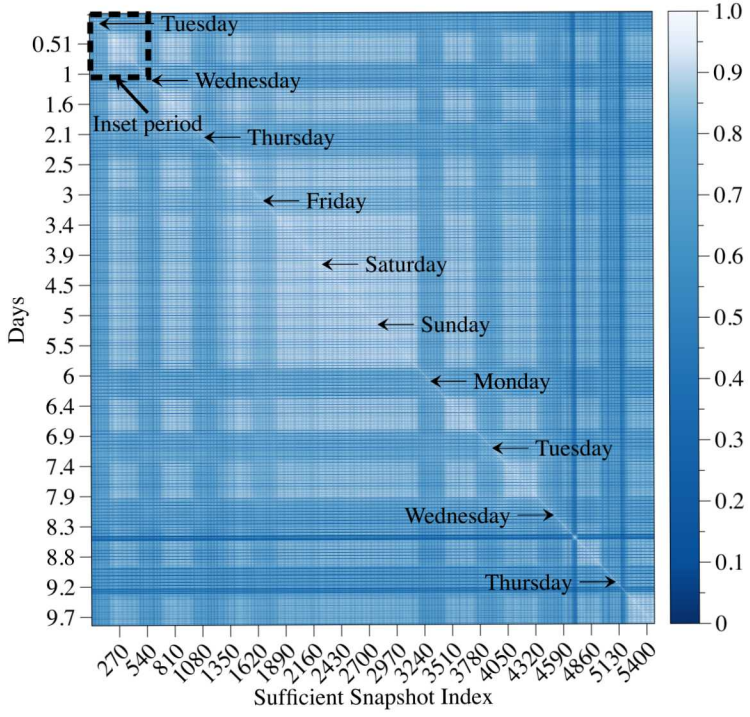


Results – LANL Netflow

- Regular structure
  - Small scale: Repeating network activity every ~70 minutes that drastically changes EA activity (repeating dark blue bands)
  - Medium scale: Daytime hours (0–7, 22-26) vs. nighttime hours (8–21)
  - Large scale: Weekdays vs. weekends



(b) LANL First Day



(c) LANL First 10 Days

Some Related Work

- Varying snapshot size
  - Studied how various graph properties change with snapshot size
  - Most consider cumulative snapshots
  - Krings, Leskovec, Rocha
- Identifying appropriate snapshots
  - Snapshots of fixed size
  - Balancing within snapshot variance with between snapshot compression (Sulo)
  - Focus on community detection (GraphScope, Sun)
  - Off-line techniques (DAPPER, Cáceres)
  - Merging neighboring snapshots (ADAGE, Soundarajan)
- Most focus on snapshots of fixed size – may not perform well when underlying data undergoes a fundamental shift
- None focus on balancing under and oversampling

Summary

- EASEE – an algorithm to appropriately build a sequence of static graphs (graph snapshots) from dynamic data
  - Avoid over and/or undersampling
  - Adaptively determines snapshot sizes
  - Forecasts future graph size
  - Possibly merges two or more snapshots into a single static graph
  - Executes in real time
  - Applied to various real and synthetic datasets
- Additional work (not included here)
  - Compared EASEE with existing algorithm ADAGE*
  - EASEE can identify "problematic" datasets – i.e., data that should be analyzed as full temporal graph
- Future work
  - Generalize synthetic data model, statistical models for EA times, densification, dying edges

* S. Soundarajan, A. Tamersoy, E. Khalil, D. Horng Chau, T. Eliassi-Rad, B. Gallagher, and K. A. Roundy. 2014. ADAGE : A Framework for Generating Adaptable Intervals from Streaming Edges.

Back up

Two Random Graph Models

- Erdős-Rényi[1] / Gilbert[2] $G(n, p)$
  - Specify the number of nodes n
  - Two nodes u, v are connected by an edge with probability p, independent of all other edges
- Stochastic blockmodel[3]
  - Specify the number of nodes n
  - Partition node set $\{1, \ldots, n\}$ into disjoint sets $c_1, \ldots, c_r$ called communities
  - $\mathbf{p}$ is an r x r matrix of probabilities, where $u \in c_i$ and $v \in c_j$ are connected by an edge with probability p_{ij}
  - If all $p_{ij} = p$, then we recover $G(n, p)$

1. P. Erdős and A. Rényi, On random graphs, Publicationes Mathematicae 6 (1959), 290– 297.

2. E. N. Gilbert, Random graphs, Annals of Mathematical Statistics 30 (1959), no. 4, 1141–1144.

3. P. W. Holland, K. B. Laskey, and S. Leinhardt, Stochastic blockmodels: First steps, Social Networks 5 (1983), no. 2, 109–137.