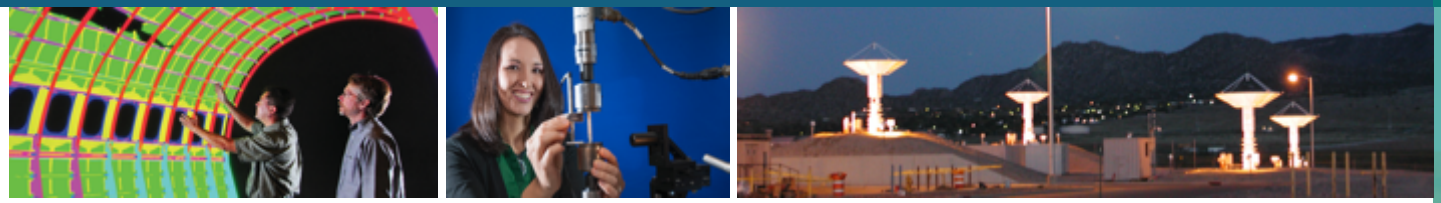
SAND2021-0569C

Sandia National Laboratories

Partitioning Communication Streams into Graph Snapshots



Jeremy D. Wendt

With Richard V. Field, Cynthia A. Phillips, Arvind Prasadan, Tegan Wilson, Sucheta Soundarajan (Syracuse U), Sanjukta Bhowmick (U of North Texas)

LAB RESEARCH TECHNICAL EXCHANGE

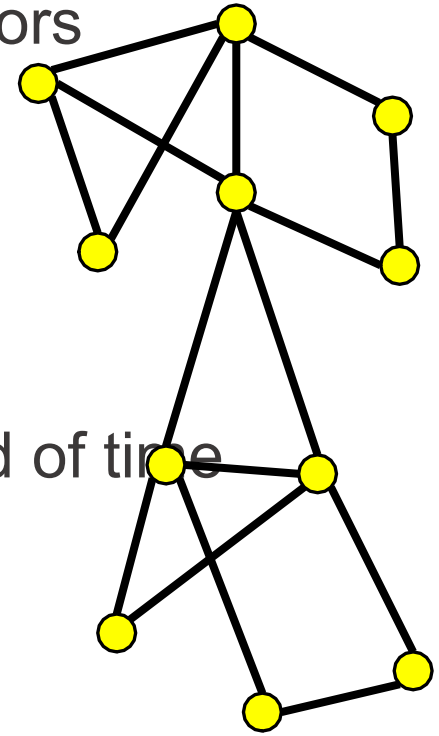
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Overview

- Graph snapshots from communication data
- EASEE
- Tests on Synthetic Datasets
- Real Datasets
- Tests and Results

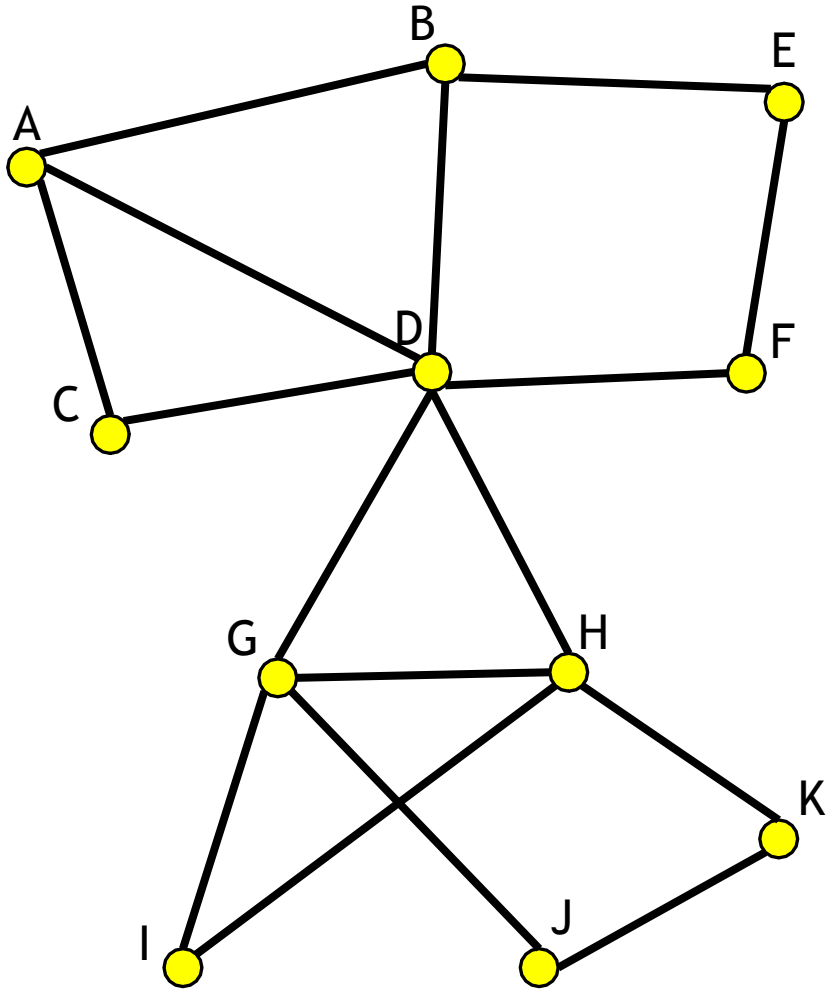
Graphs from Communication Data

- Many graphs can be *crawled* – request a node identify its neighbors
  - E.g., web crawls, questionnaires
- Communication data *advertises* edges at data-specified times
  - E.g., emails, computer networks
  - Call these *edge advertisements* (EAs)
- Thus, communication-data graphs require sampling over a period of time
  - Underlying system may change drastically during this time period



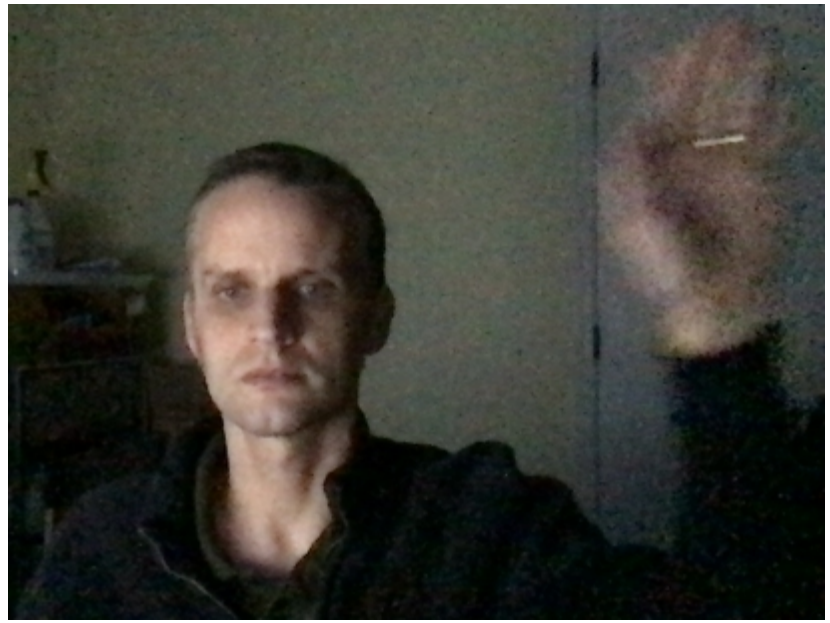
Graph Building Example

A, B, 3
A, C, 5
D, G, 9
E, F, 9
H, G, 10
J, G, 12
J, K, 12
A, B, 17
A, D, 17
C, D, 18
D, F, 21
…



Photography comparison

- Collect photons for some period of time
  - Photon rate (e.g., light strength)
  - Scene motion (e.g., waving hand)
- Collect edge advertisements for some period of time
  - Advertisement rate (e.g., communication frequency)
  - System motion (e.g., people join/leave, relationships form/decay)
- Determine proportion sampled vs. scene stability



Real-world Need

- Analytics run against graph data generated from EAs
- Graph from too little data will miss important connections
  - Graphs may appear disjoint when actually well connected
- Graph from too much data (esp. across system changes) will show edge relationships when edges never existed within similar periods
  - Nodes may appear far closer related than they truly are
- Decisions made from analytics on such datasets may be wrong
- Graphs formed as the data streams is best – decisions can be more timely

- **Leskovec2007**: The more temporal data you add, the denser the graph

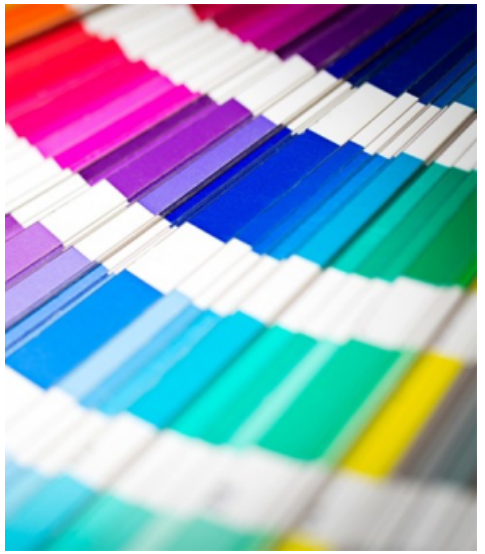
Previous Work

- **Sun2007**: An off-line technique for trading off community change with number of sampling windows
- **Sulo2010**: An off-line technique for trading off within-window variance and between-window compression
- **Caceres2013**: An off-line technique for identifying optimal window sizes
- **Soundarajan2014**: An on-line technique for identifying when network measures converge

Edge Advertisements into Snapshots using Evolving Expectations

- EASEE has two key steps:
  - Sufficient Window Detection
    - Minimal data for a graph collected
  - Neighboring Window Merging
    - Merge neighboring sufficient snapshots if system motion was slow enough

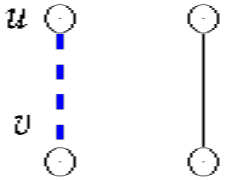
- EASEE benefits
  - Few parameters with good default settings for most
  - Executes in real-time with EAs
  - Identify problematic datasets
  - Adaptively identify snapshot intervals
  - Predict future graph sizes



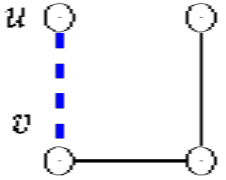
Definitions

- Edge advertisement (u, v, t)
  - Communication takes place between entities u and v at time t
  - We assume a streaming list of EAs
  - An edge may advertise more than once
- *Type N2*: New edge with 2 new nodes
- *Type N1*: New edge with 1 new node
- *Type N0*: New edge with 0 new nodes
- *Type R*: Repeat edge

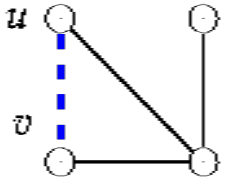
N2-type



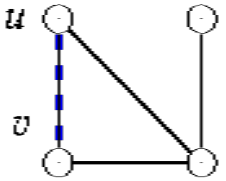
N1-type



N0-type

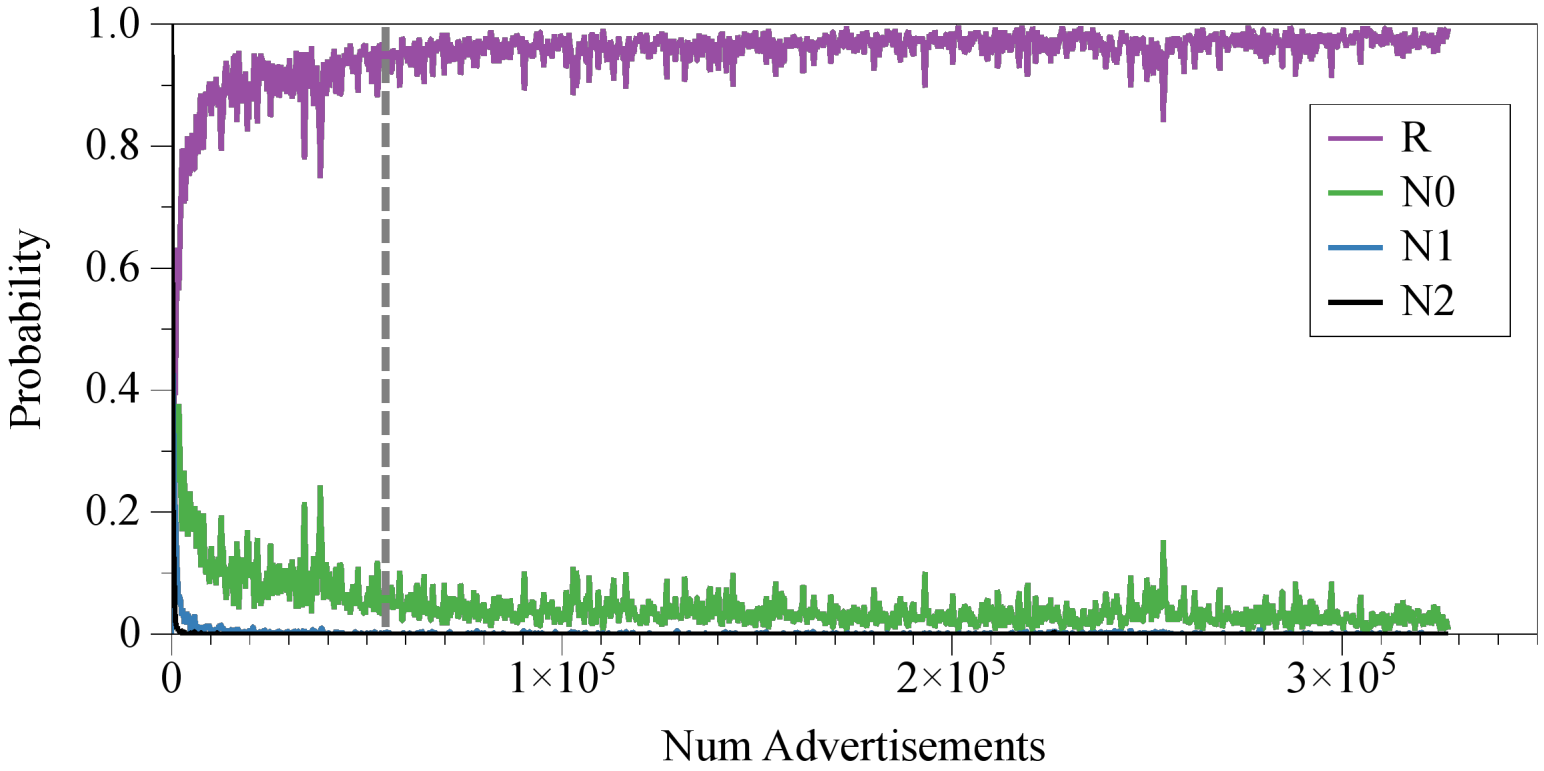


R-type



Estimate probabilities of these types

- Given the past *k* EAs, compute the percentage of those that belong to each type



Predict future graph size

- Our most important equations:

$$\mathbb{E}[m_{x+y} - m_x] = y\left(1 - p(\mathrm{R}, x)\right)$$
$$\mathbb{E}[n_{x+y} - n_x] = y\left(p(\mathrm{N1}, x) + 2\,p(\mathrm{N2}, x)\right)$$

  - where *x* is the number of EAs already seen, *y* is number of EAs more to see, *m* is number of edges, *n* is number of nodes, and *p*(T, *x*) is the current probability of EA type T after *x* EAs.

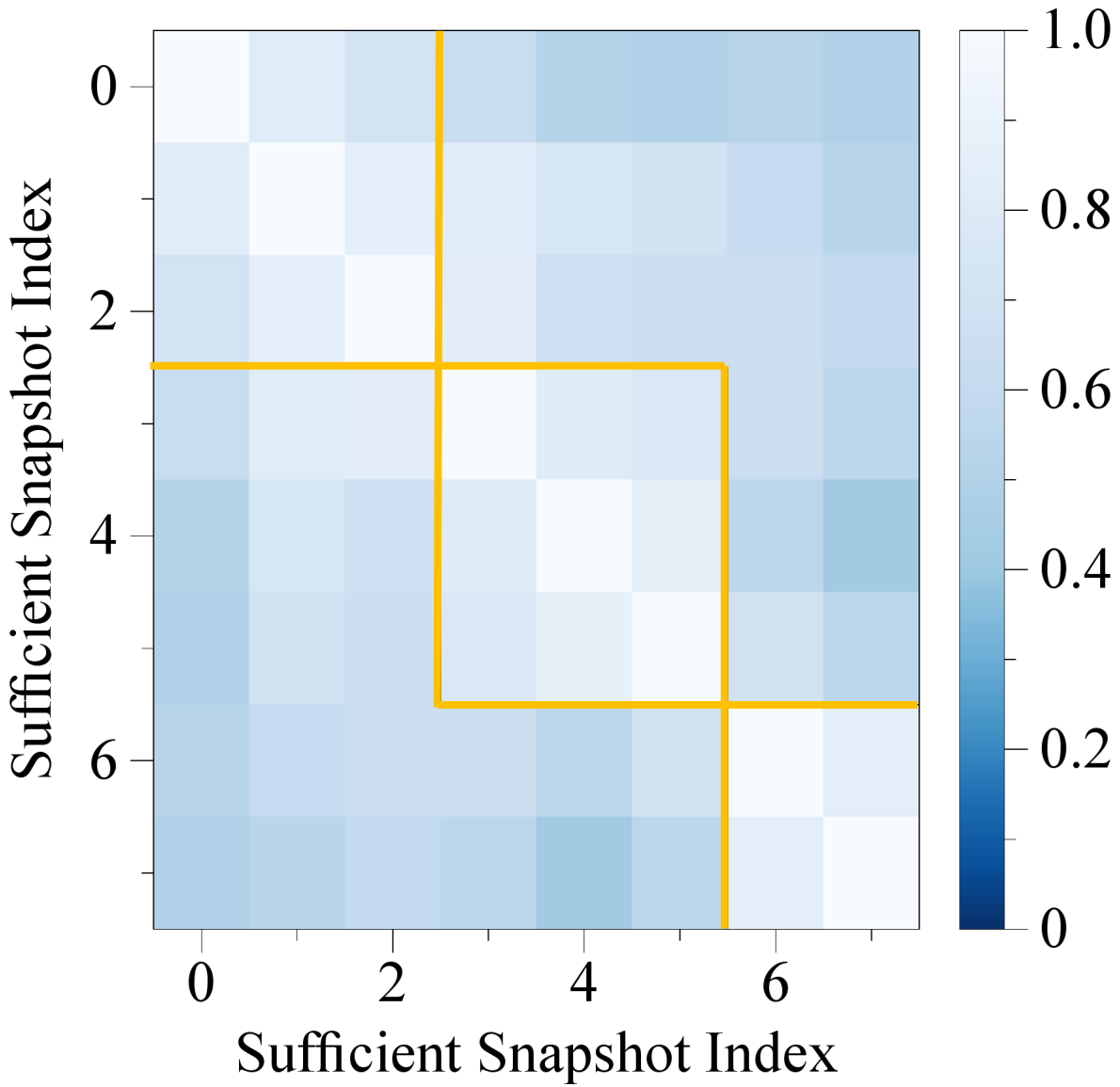
- Essentially, this allows us to predict the most expected increase in number of edges (*m*) and number of nodes (*n*) after *y* more EAs.
  - We can then collect *y* more EAs and monitor the error in these predictions
  - We can also watch these predictions and look for convergence or drastic changes in the prediction

- Create “sufficient sample” snapshots when convergence or drastic increase in predictions

Step 2: Consider merging

- Given a stream of sufficient snapshots, can we merge neighboring snapshots?
  - Why merge? A *minimum* snapshot may not have all of the data you want for your downstream analytic.
  - Why not merge? If two neighboring minimum snapshots are very different, merging would result in a “system blurred” graph snapshot
- How to measure if snapshots are different?
  - Sufficient snapshots preserve EA count per edge… compute cosine similarity of neighboring snapshots’ EA counts per edge

Cosine Similarities Example – EU Emails



Synthetic Dataset Test 1: How important to detect changes?

- Given a stream of EAs on the a set of nodes, at some time *t*, the communities drastically change
  - What happens if you miss the change time *t* for later graph analysis?
  - Nerd details: SBMs with 10 blocks, $p_{in0} = 0.05$, $p_{in1} = 0.04$, and $p_{out} = 0.005$
- Test ran 20 times, average scores shown next slide
  - Nerd details: Louvain-identified communities tested against SBM-true communities using AMI
- High community detection scores means that EASEE would do a good job on your community detection problems
  - Very low scores means your identified communities are essentially noise

Synthetic Dataset Test 1: How important to detect changes?

Name	First EA	Last EA	Matched to	AMI Mean	AMI Stdev
Ground Truth	0	49,999	G_0	0.948	0.009
Ground Truth	50,000	99,999	G_1	0.708	0.044
EASEE	0	50,271.4	G_0	0.947	0.008
EASEE	50,272.4	99,999	G_1	0.704	0.038
Fixed Split 0.2	10,000	59,999	G_0	0.775	0.027
Fixed Split 0.4	20,000	69,999	G_0	0.096	0.011
Fixed Split 0.6	30,000	79,999	G_1	0.029	0.003
Fixed Split 0.8	40,000	89,999	G_1	0.094	0.014

Best possible scores

EASEE scores

Error scores

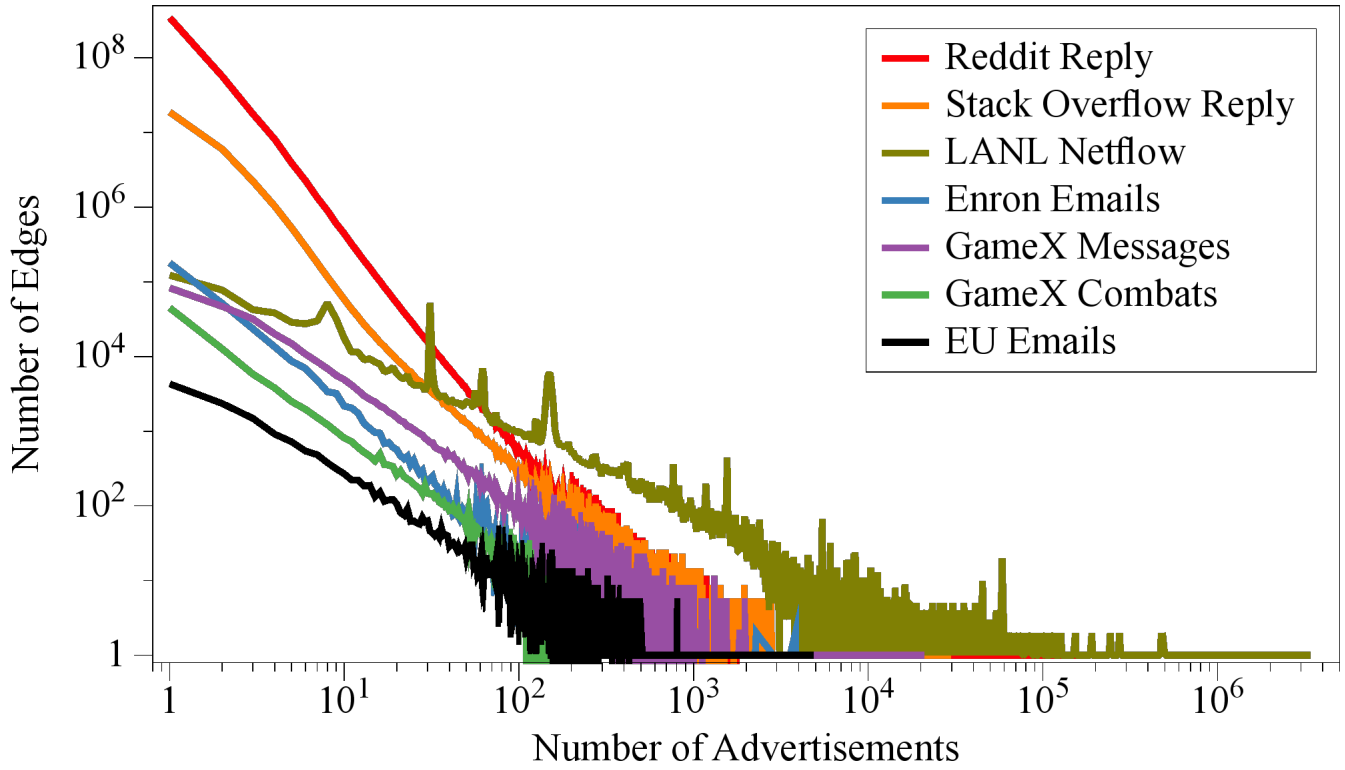
Other tests performed (paper has details)

- How do EA change rate, timing of change, and amount of change affect change detection?
  - Slower changes and smaller changes harder to detect
  - Changes very recently after snapshot began can be impossible to detect
  - Previous techniques completely ignore the change – only line up with change by random chance
- When creating a stream of snapshots, how does start time affect resulting snapshots?
  - EASEE generally converges to the same change moment (exactly!) within 10 snapshots on a variety of datsets

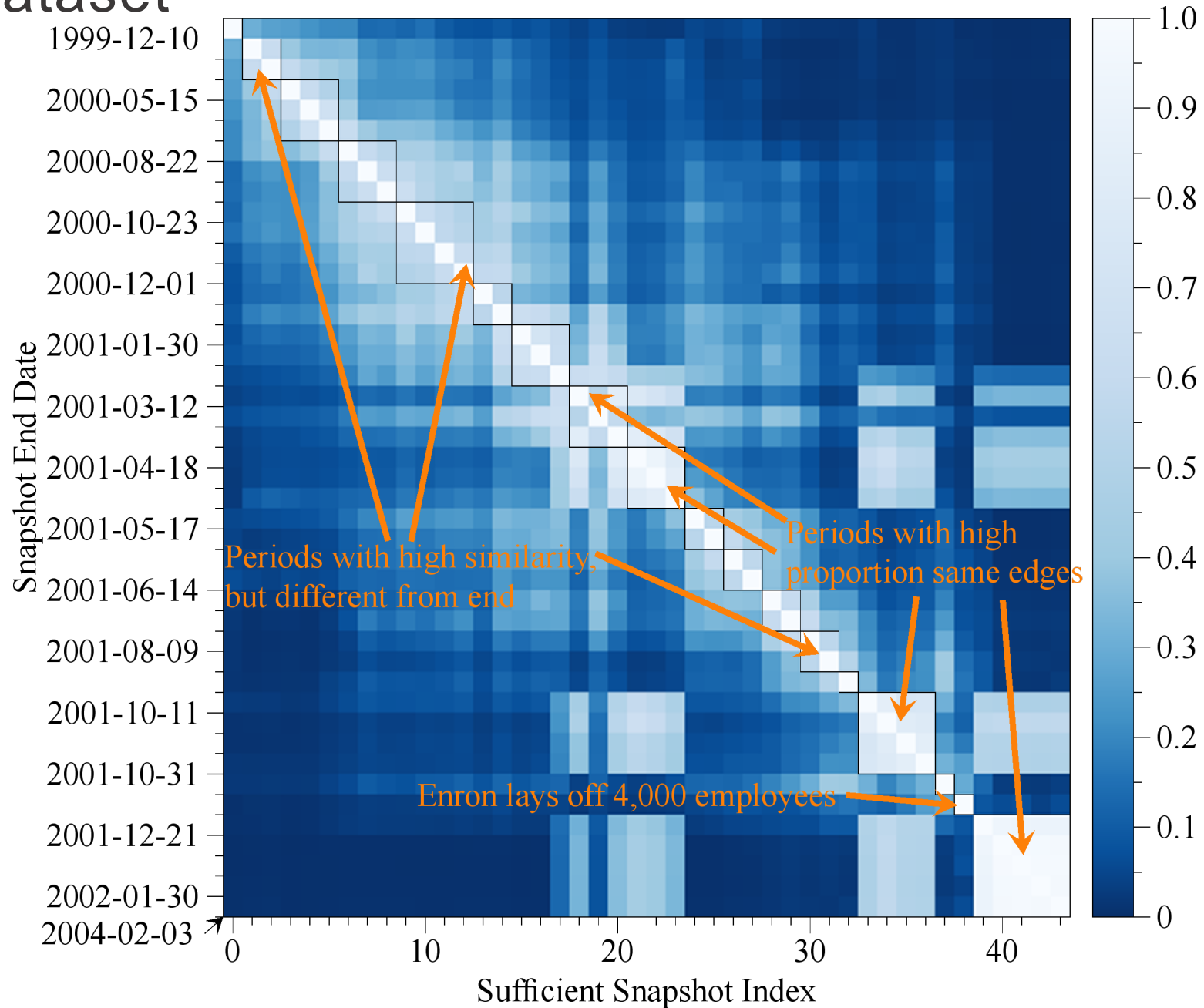
Real-world Datasets

- *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
- *Reddit Reply*: Log of posts and comments on social news
- *LANL Netflow*: 32 days of computer network traffic (both human and automated)

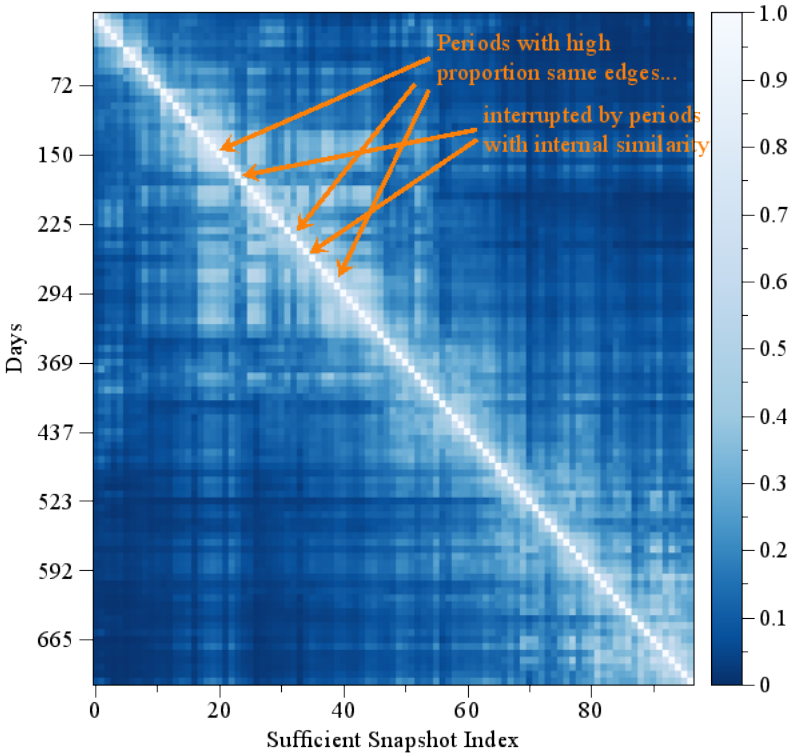
Name	Number of EAs	Number of Nodes	Number of Edges
EU Core Emails	327,336	986	16,064
Enron Emails	1,283,755	84,511	316,061
GameX Combats	500,327	10,589	86,351
GameX Messages	4,515,396	22,442	293,860
Stack Overflow Reply	63,496,479	2,601,977	29,541,284
Reddit Reply	646,024,723	8,901,033	437,747,667
LANL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



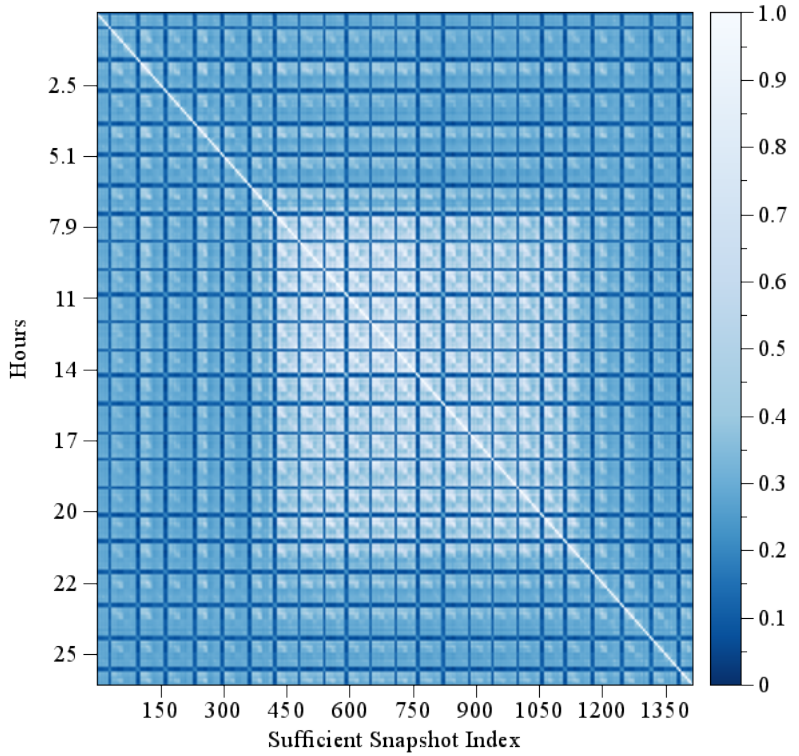
Enron Dataset



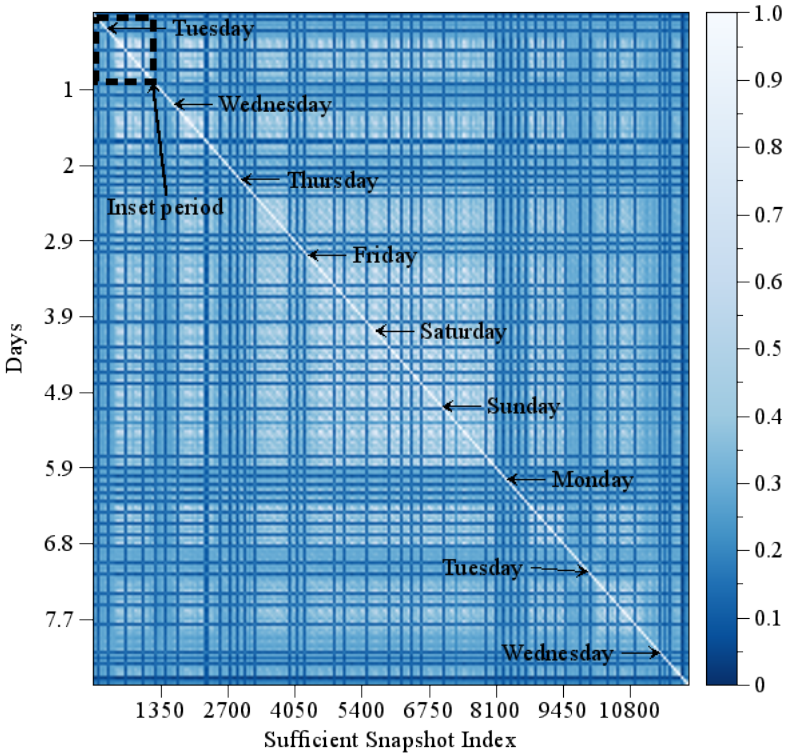
Other Datasets 1



Game Messages



Network Traffic – 1 day



Network Traffic – 1 week

Other Datasets 2

- GameX Combats and Reddit both *never* merged even at relatively low thresholds
  - Indicates that the edges are shifting very quickly in these datasets – faster than the EA rate can support
  - Forming static graphs from such datasets may be unwise!

Conclusion/Questions

- EASEE detects snapshots in changing data … and merge only when changes are sufficiently small
- EASEE monitors densification and differentiates between “converging” densification and “changing” densification
- EASEE can detect when the underlying graph shifts and reverts between states

- Questions?
- jdwendt@sandia.gov