Detecting Anomalous Activity in Computer and Phone Networks

Brian Thompson James Abello

RUTGERS

Outline

- Introduction and Motivation
- Model and Approach
- The MCD Algorithm
- Experimental Results
- Conclusions and Future Work

Outline

- Introduction and Motivation
- Model and Approach
- The MCD Algorithm
- Experimental Results
- Conclusions and Future Work

Communication Networks

- People like to talk

phone calls

email

Twitter

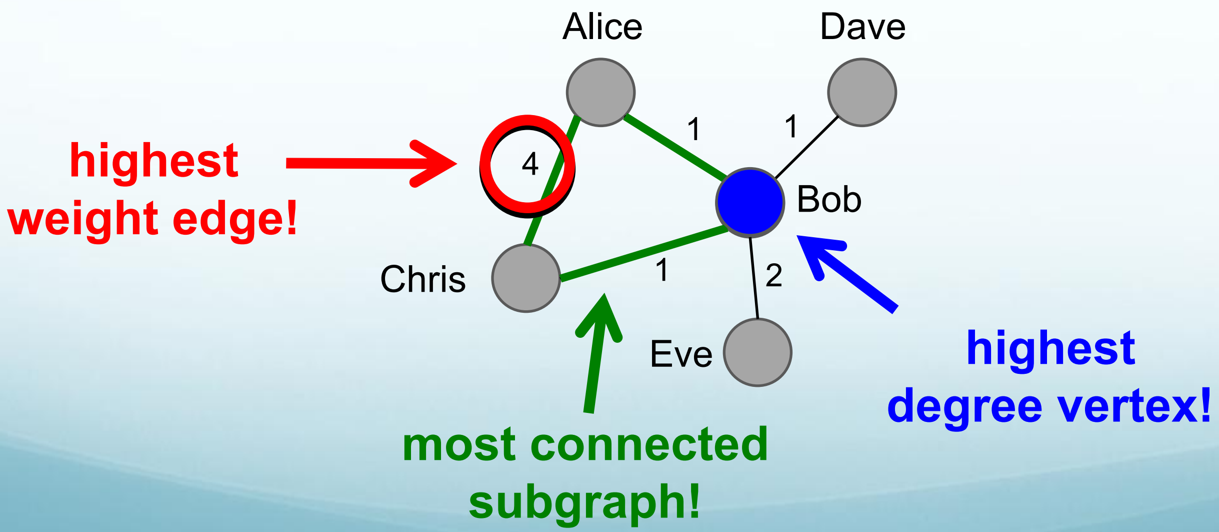
IP traffic

- May be stored in communication logs:

Sender	Receiver	Timestamp
Alice	Bob	Nov. 16, 2010 5:20pm
Alice	Chris	Nov. 17, 2010 9:45am
Bob	Dave	Nov. 17, 2010 7:00pm

Communication Networks

- Commonly visualized as graphs, where nodes are people and edges signify communication
- Edge weights may indicate the quantity or rate of communication
- Graph analysis tools may then be applied to gain insights into network structure or identify outliers



Motivation

- Communication networks are **BIG** and *highly volatile*
- Traditional techniques are ineffective for analyzing network dynamics, dealing with lots of streaming data
- We address questions of temporal and structural nature

Traditional Question	Our Question
Which nodes have the highest degree?	Which nodes have seen a recent change in connectivity patterns?
Which subgraphs are the most well-connected?	Which subgraphs have shown a sudden increase in activity?

Applications

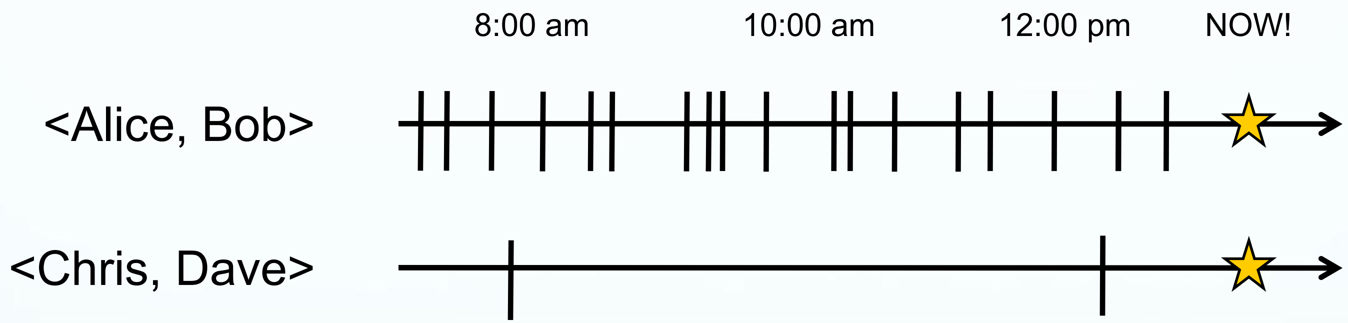
- Monitor suspected malicious individuals or groups
- Detect the spread of viruses on a local network
- Identify blog posts that have achieved sudden popularity among a small subset of users, that might otherwise fall below the radar
- Flag suspicious email or calling patterns without examining the content or recording conversations

Outline

- Introduction and Motivation
- Model and Approach
- The MCD Algorithm
- Experimental Results
- Conclusions and Future Work

Recency

- Our goal is to detect anomalous activity *as it is happening*.
- Key idea: **more recent = more relevant**
- For each pair of people, how recently did they communicate?

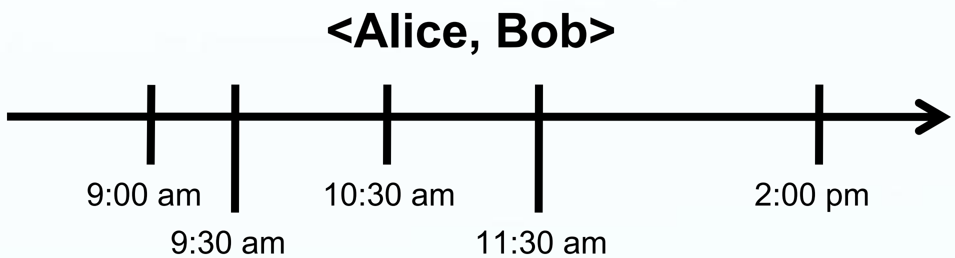


- Problem: The most frequent communicators will always seem “recent”, overshadowing others’ behavior.

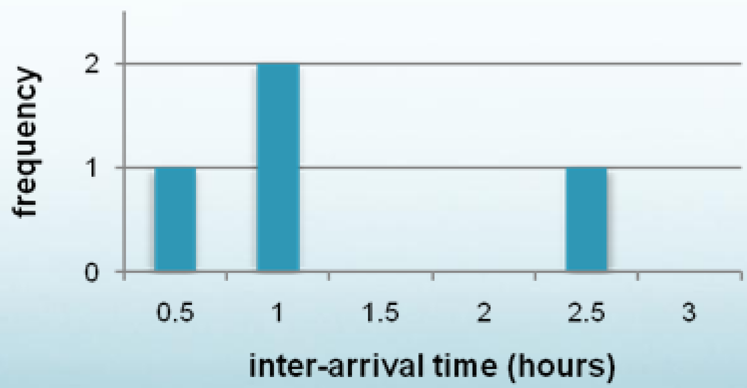
We call this ***time scale bias***.

An Edge Model

- We model communication across an edge as a *renewal process*: a sequence of time-stamped events

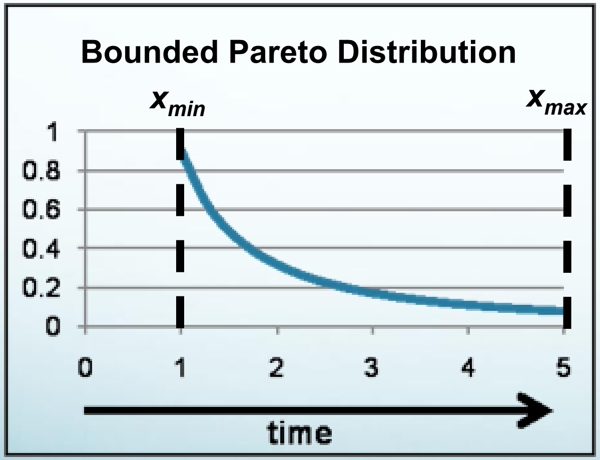


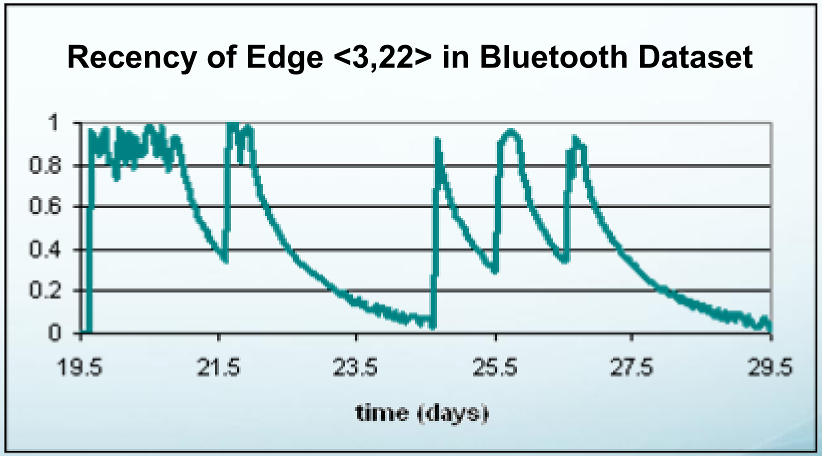
- sampled from a distribution of inter-arrival times (IATs)



Recency

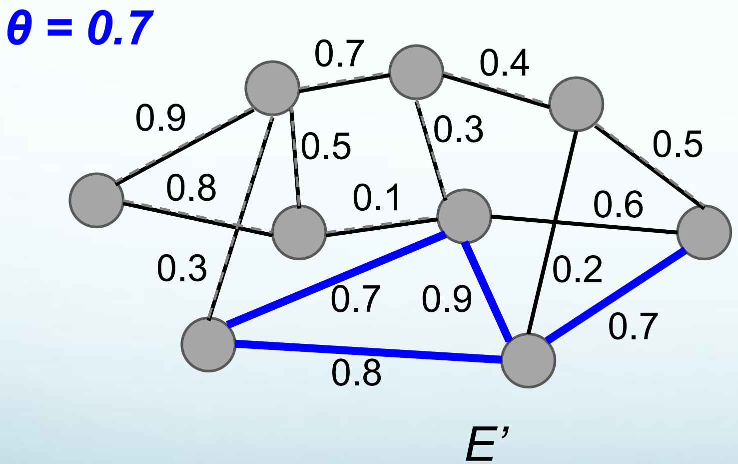
- The recency function **Rec : 2^T x T → [0,1]** assigns a weight to an edge ***e*** at time ***t*** based on the age of the renewal process (time since the last event), decreasing from *age* = **0** to $\boldsymbol{x_{max}}$.
- Given an IAT distribution, there is a unique such function that is uniform over [0,1] when sampled uniformly in time.
- This property eliminates time scale bias.





Divergence

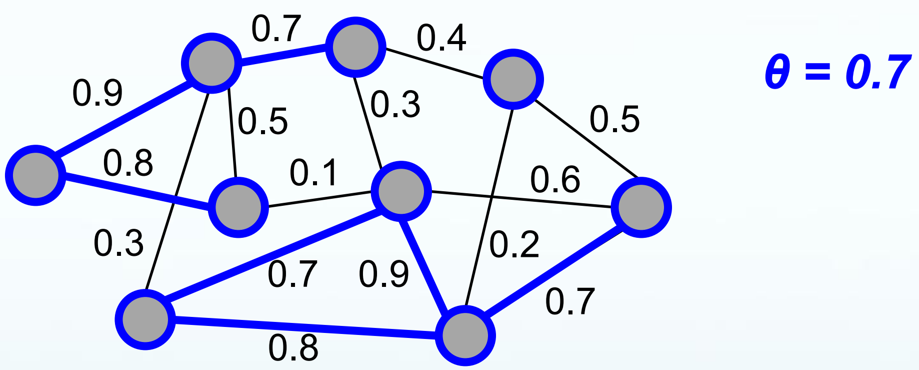
- Consider the weighted graph *G = (V,E)* representing a communication network, with *w(e) = Rec(e)*.
- For $E' \subseteq E$, let $X_{E',\theta}$ = *# of edges in E' with Rec(e) ≥ θ*.
  We define $Div_\theta(E') = \dfrac{1}{P(X \geq X_{E',\theta})}$, where $X \sim Bin(|E'|, 1-\theta)$.



- $|E'| = 6$
- $X_{E',0.7} = 4$
- $P(X \geq 4) = 0.07$
- $Div_{0.7}(E') = 14.19$

Maximal Components

- A (maximal) θ-component of $G = (V,E)$ is a connected subgraph $C = (V',E')$ such that
  1. $w(e) \geq \theta$ for all e in E'
  2. $w(e) < \theta$ for all e not in E' incident to V'



- The set of θ-components form a partition of V, for all θ in [0,1].

Algorithmic Challenges

- Combinatorial explosion: There are too many possible subgraphs – $2^{|E(G)|}$ to be exact!
  - Only look at connected components
- There are still too many possible subgraphs!
  - Only look at maximal θ-components
- How do we know what's the right θ threshold?
  - Try them all!

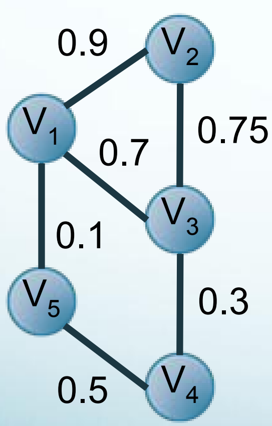
We now present the MCD (Maximal Component Divergence) Algorithm.

Outline

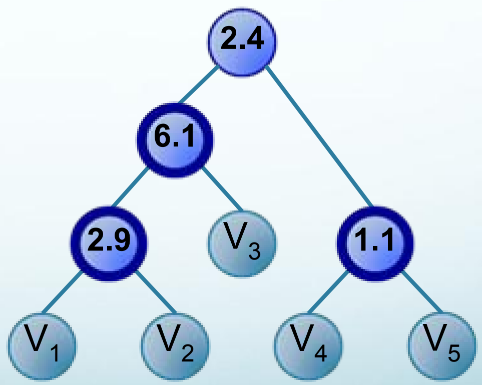
- Introduction and Motivation
- Model and Approach
- The MCD Algorithm
- Experimental Results
- Conclusions and Future Work

The MCD Algorithm

1. Calculate edge weights using the Recency function
2. Gradually decrease the threshold, updating components and divergence values as necessary
3. Output: Disjoint components with max divergence

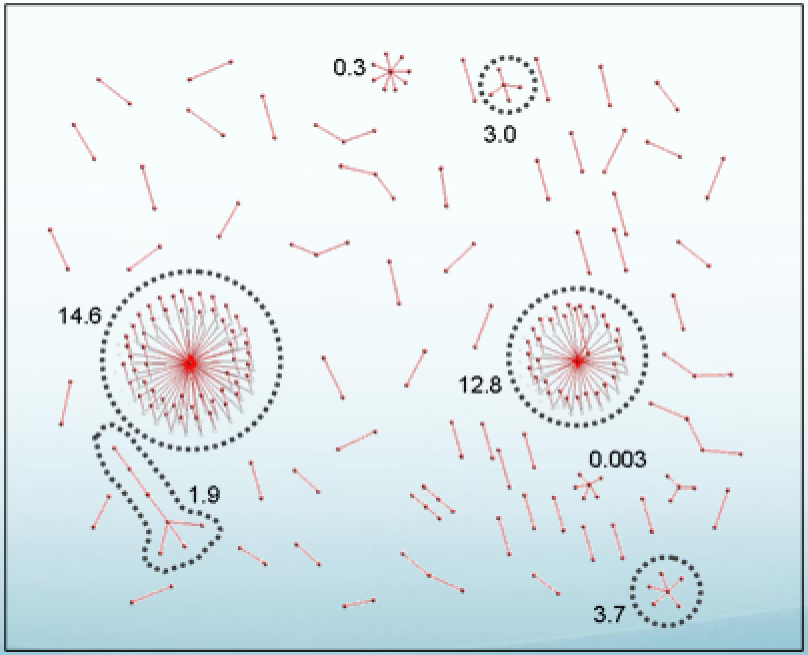


θ	Component	Div(C)
0.9	$\{V_1,V_2\}$	2.908
0.75	$\{V_1,V_2,V_3\}$	2.723
0.7	$\{V_1,V_2,V_3\}$	6.132
0.5	$\{V_4,V_5\}$	1.143
0.3	$\{V_1,V_2,V_3,V_4,V_5\}$	2.380
0.1	$\{V_1,V_2,V_3,V_4,V_5\}$	1.882



Sample Output

MCD	θ	#V(C)	E-frac	%E(C)	%E(G)
14.57	0.07	54	53/212	0.25	0.08
12.84	0.08	32	31/88	0.35	0.08
3.70	0.10	6	5/7	0.71	0.10
2.97	0.18	5	4/4	1.00	0.14
1.91	0.05	7	6/41	0.15	0.04



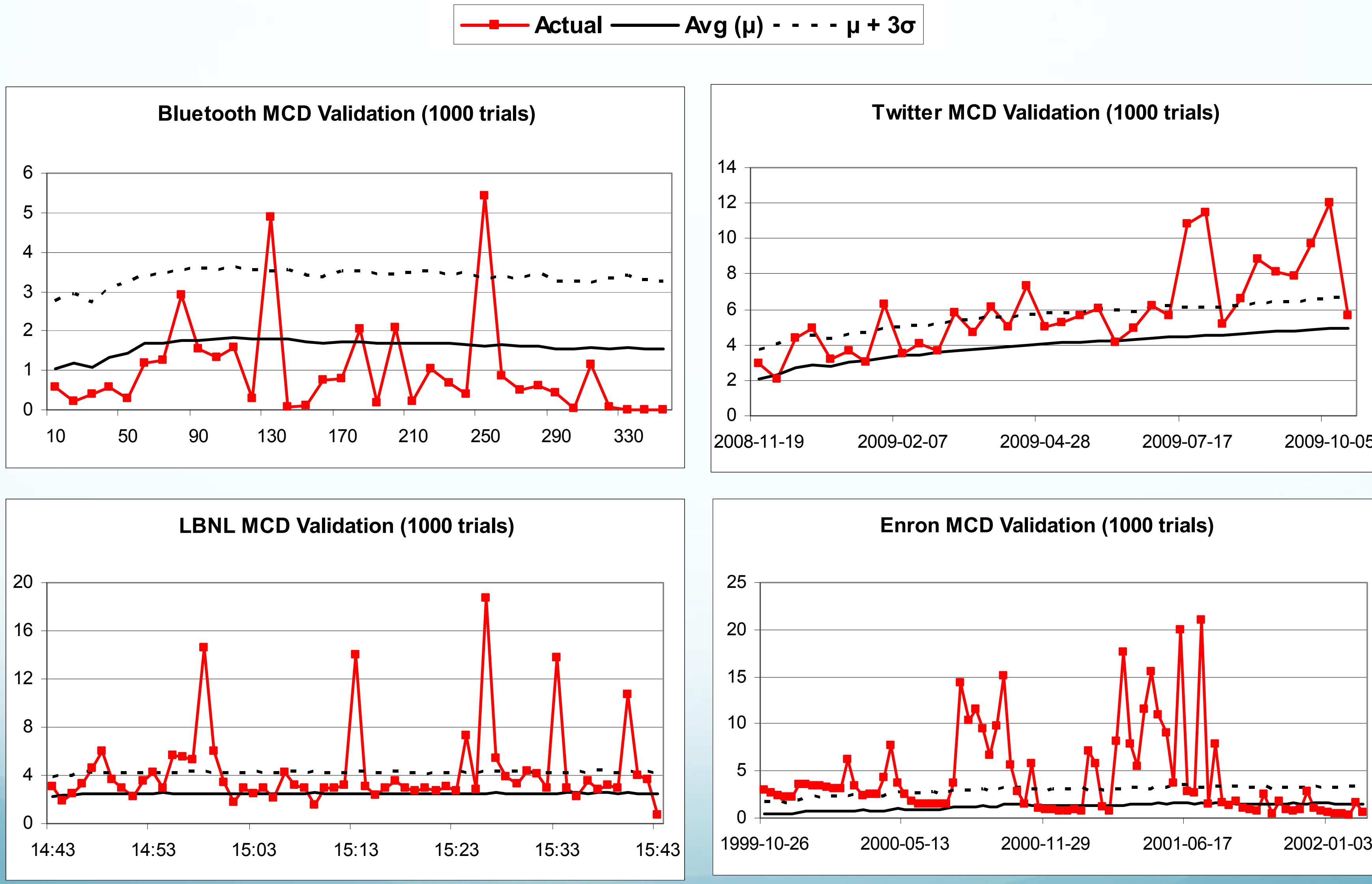
Outline

- Introduction and Motivation
- Model and Approach
- The MCD Algorithm
- Experimental Results
- Conclusions and Future Work

Data

Dataset	# Nodes	# Edges	# Timestamps
ENRON – email network of Enron employees	1141	2017	4847
BLUETOOTH – proximity of mobile devices	101	2815	102563
LBNL – logs of IP traffic	3317	9637	9258309
TWITTER – directed messages	262932	307816	1134722

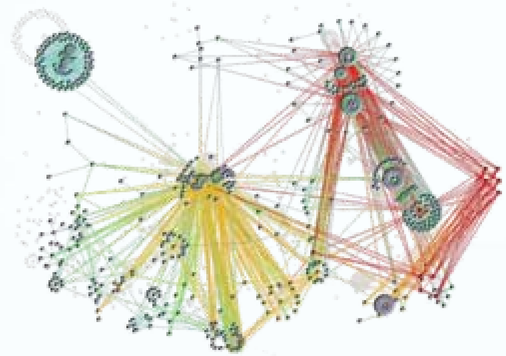
Validation of Results

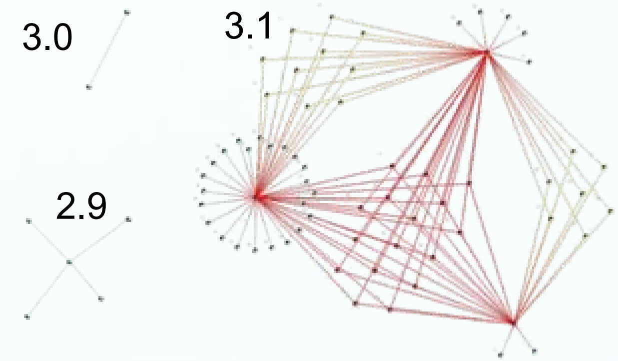


LBNL Case Study

Time:

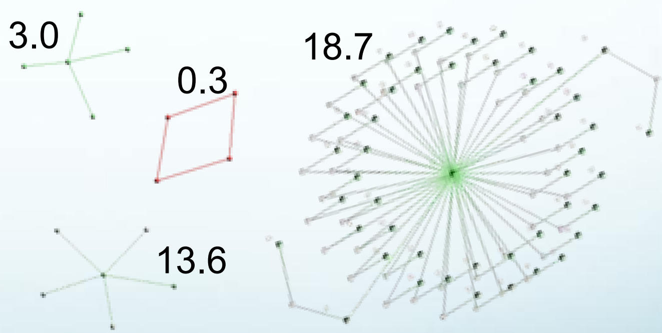
12:07pm





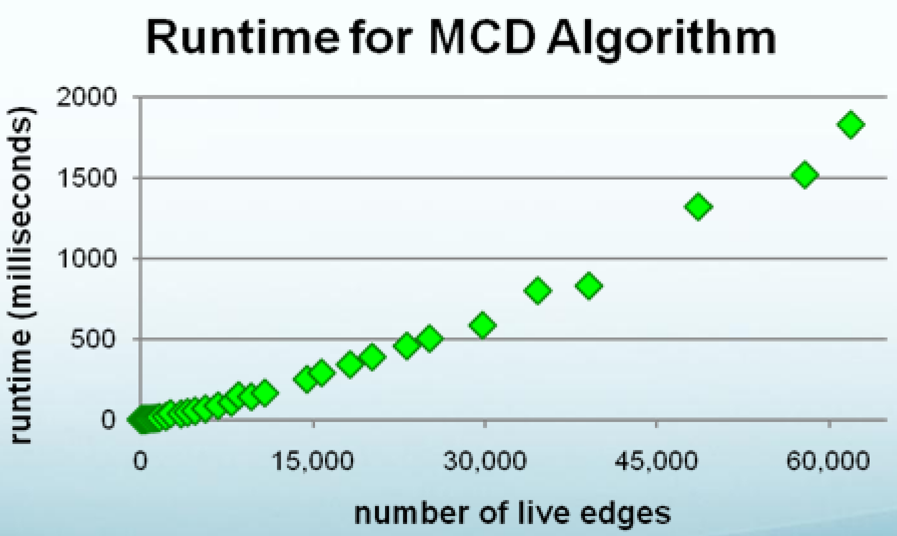
Time:

12:26pm



Complexity Analysis

- Dataset: Twitter messages, Nov. 2008 – Oct. 2009 (263k nodes, 308k edges, 1.1 million timestamps)
- Updates *O(1)* per communication
- MCD Algorithm *O(m* log *m)*, where *m* = # of edges; can be approximated in effectively *O(m)* time



Outline

- Introduction and Motivation
- Model and Approach
- The MCD Algorithm
- Experimental Results
- Conclusions and Future Work

Future Work

- Incorporate duration of communication and other edge properties into our model
- Extend our method to accommodate other data types, such as recommendation systems or hypergraphs
- Develop techniques to take past correlation of edges into account (to avoid recurring “anomalies”)
- Make it even more efficient – linear in number of nodes?

Acknowledgements

- Part of this work was conducted at Lawrence Livermore National Laboratory, under the guidance of Tina Eliassi-Rad.
- This project is partially supported by a DHS Career Development Grant, under the auspices of CCICADA, a DHS Center of Excellence.

Lawrence Livermore National Laboratory

CCICADA

Command, Control, and Interoperability Center for Advanced Data Analysis

Questions?