Fusion of Content and Context in Human Language Technology

Allen Gorin

Human Language Technology Research
National Security Agency
Fort Meade, Maryland

Graph Exploitation Workshop, August 2011

Report Documentation Page			*Form Approved* *OMB No. 0704-0188*
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.			
1. REPORT DATE **AUG 2011**	2. REPORT TYPE **N/A**		3. DATES COVERED **-**
4. TITLE AND SUBTITLE **Fusion of Content and Context in Human Language Technology**			5a. CONTRACT NUMBER
			5b. GRANT NUMBER
			5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)			5d. PROJECT NUMBER
			5e. TASK NUMBER
			5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Human Language Technology Research Na4onal Security Agency Fort Meade, Maryland**			8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)			10. SPONSOR/MONITOR’S ACRONYM(S)
			11. SPONSOR/MONITOR’S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release, distribution unlimited**			
13. SUPPLEMENTARY NOTES **See also ADA578519. Graph Exploitation Symposium. Held in Lexington, Massachusetts on August 9-10, 2011. ESC-TR-2010-094**			
14. ABSTRACT			
15. SUBJECT TERMS			
16. SECURITY CLASSIFICATION OF:		17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**SAR**	**52**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Collaborators

Carey Priebe (JHU)

John Grothendieck (BBN)

Glen Coppersmith (JHU HLT COE)

Walt Andrews (BBN)

John Conroy (IDA CCS)

Richard Cox (COE)

Mike Decerbo (BBN)

Nam Lee (COE)

Dave Marchette (NSWC)

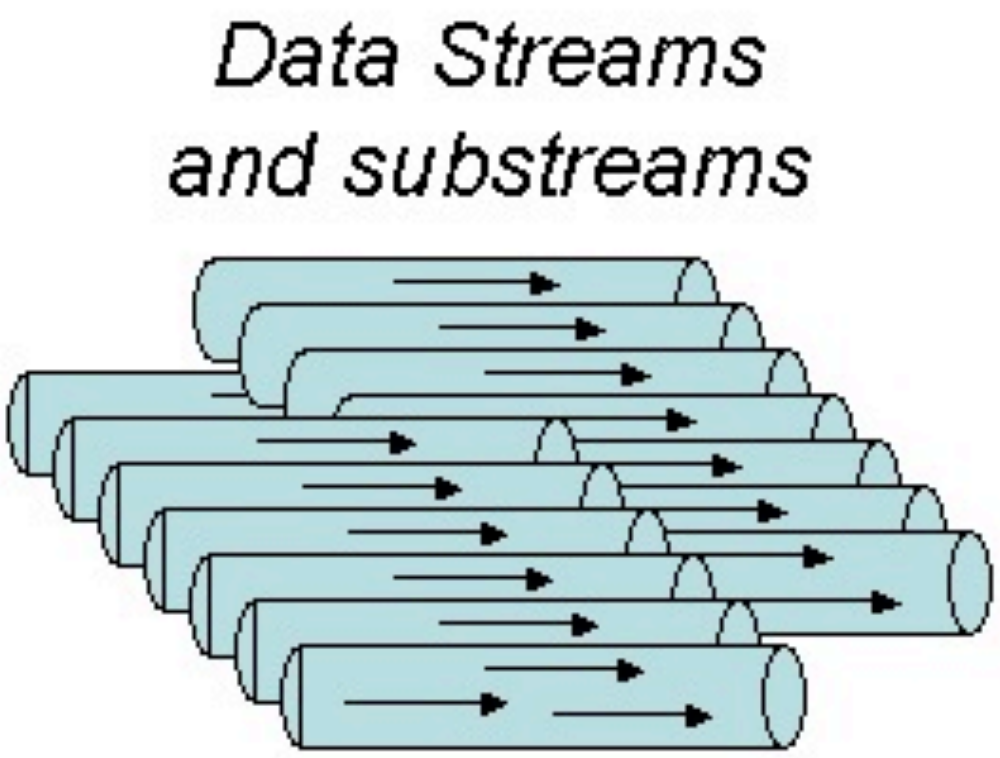
Alan McCree (MIT LL)

Youngser Park (COE)

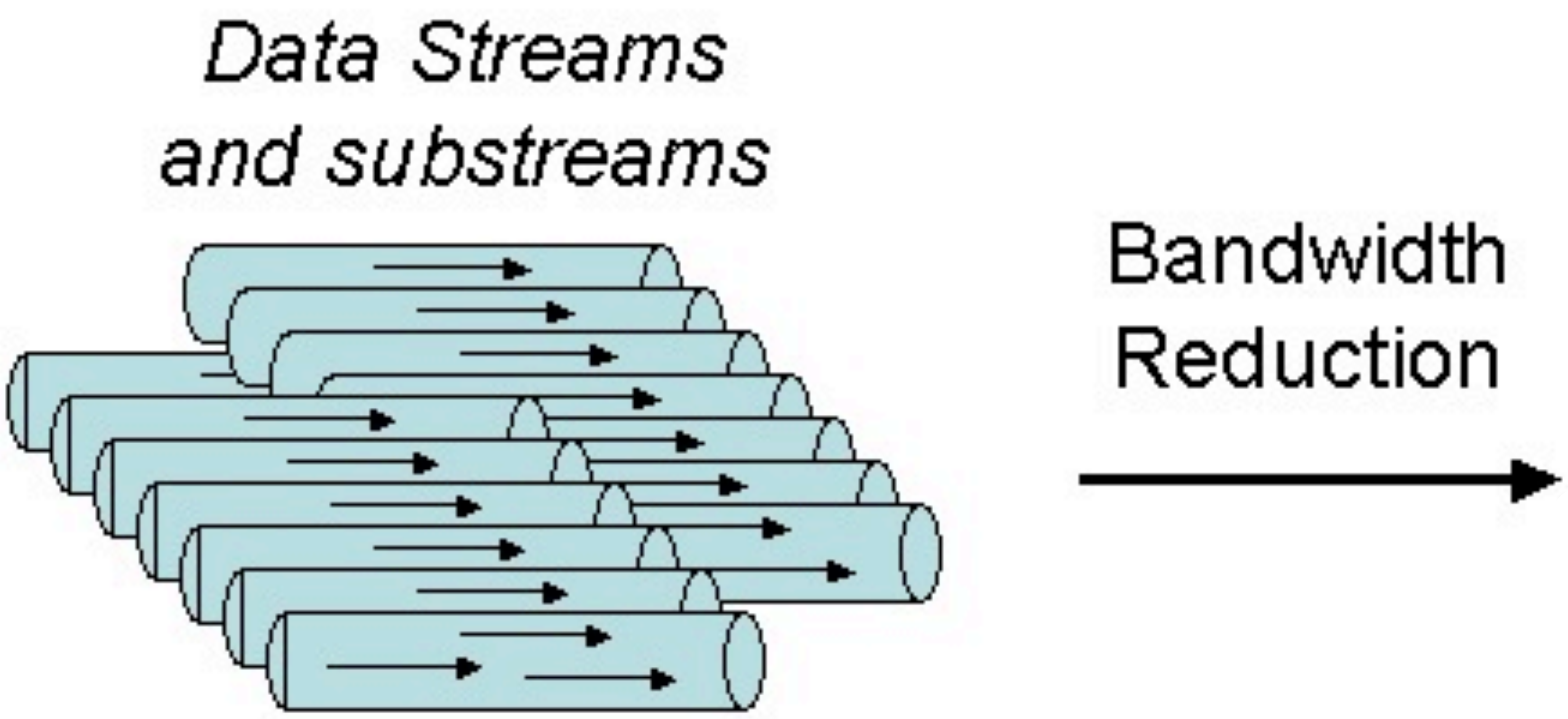
Outline

- Motivation: Coping with Information Overload
- Examples of Context and Content
- Random Attributed Graphs
- Three Tasks
  - Stream Characterization
  - *Vertex Nomination*
  - Dyadic Priors

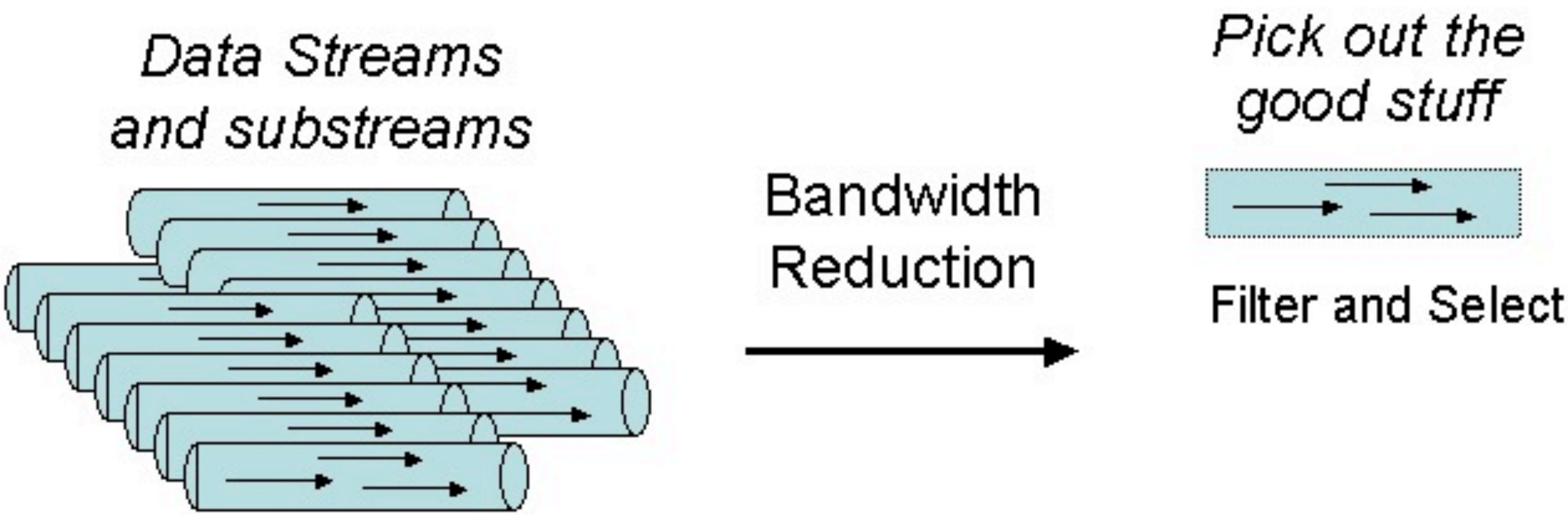
Coping with Information Overload



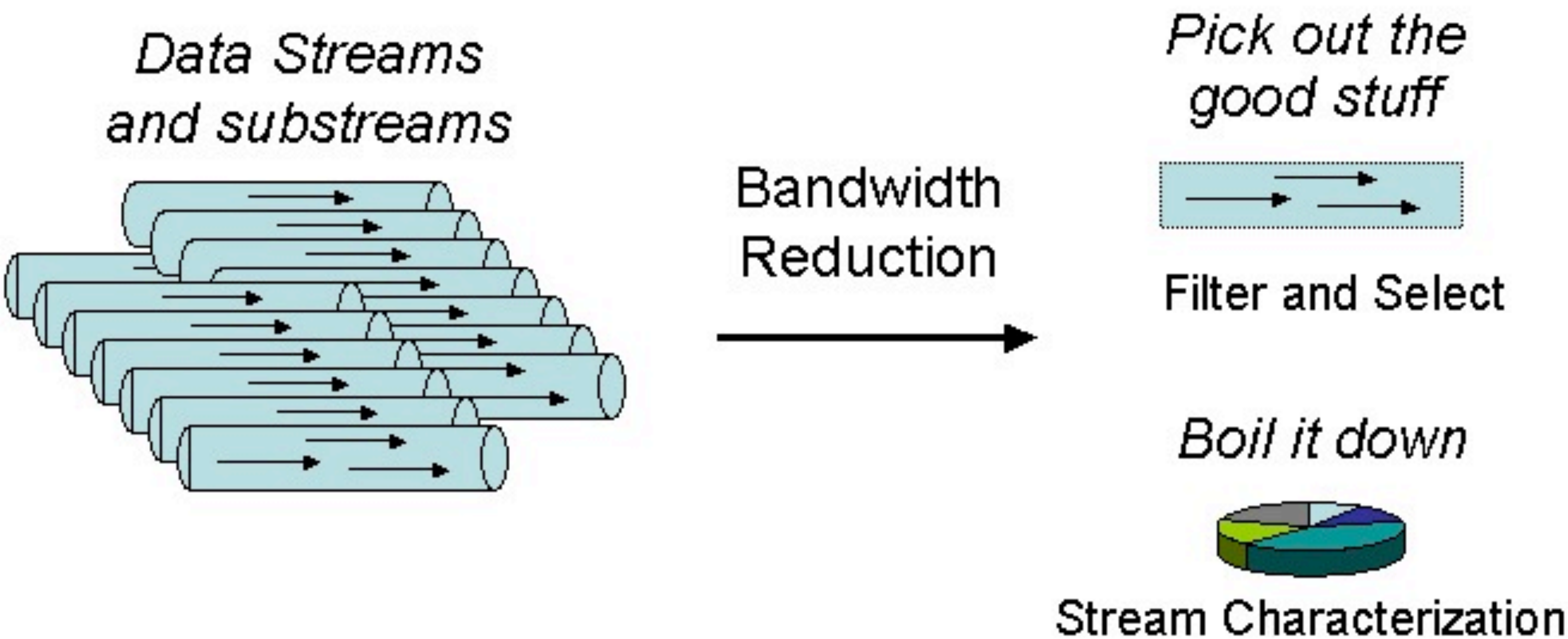
Coping with Information Overload



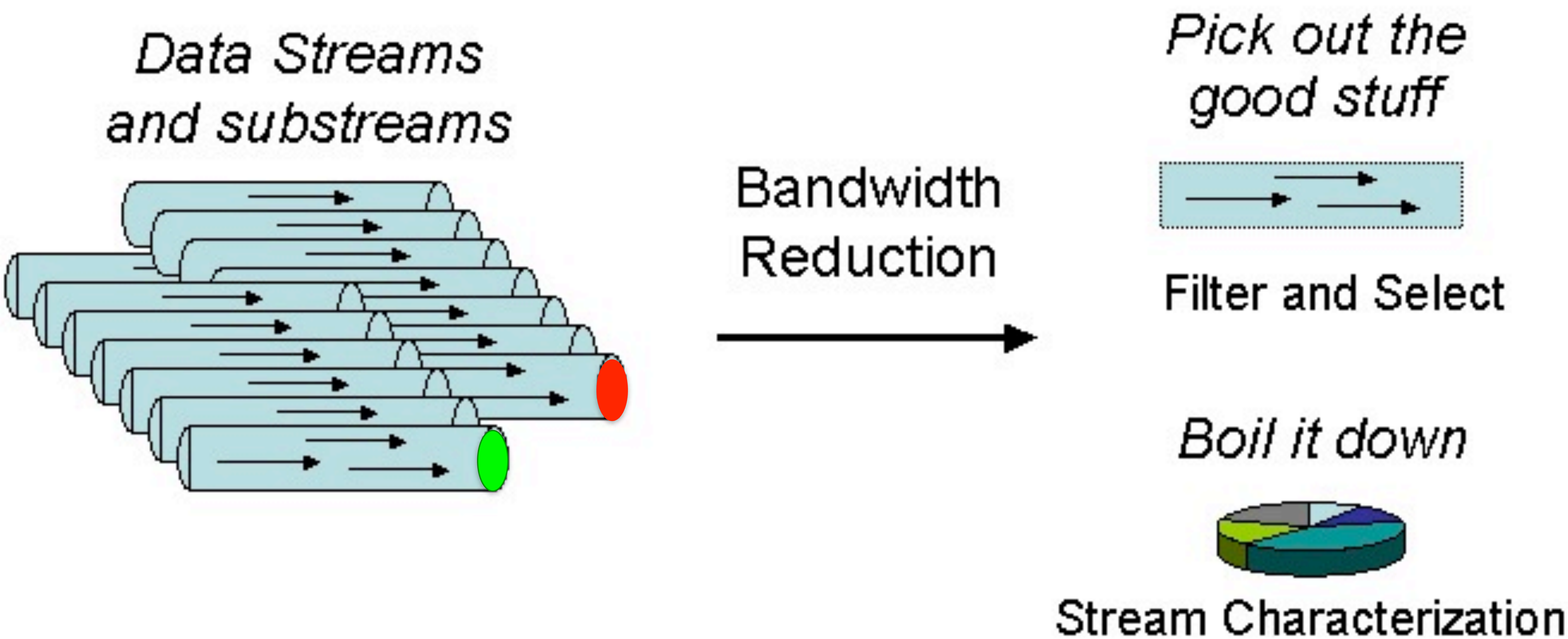
Coping with Information Overload



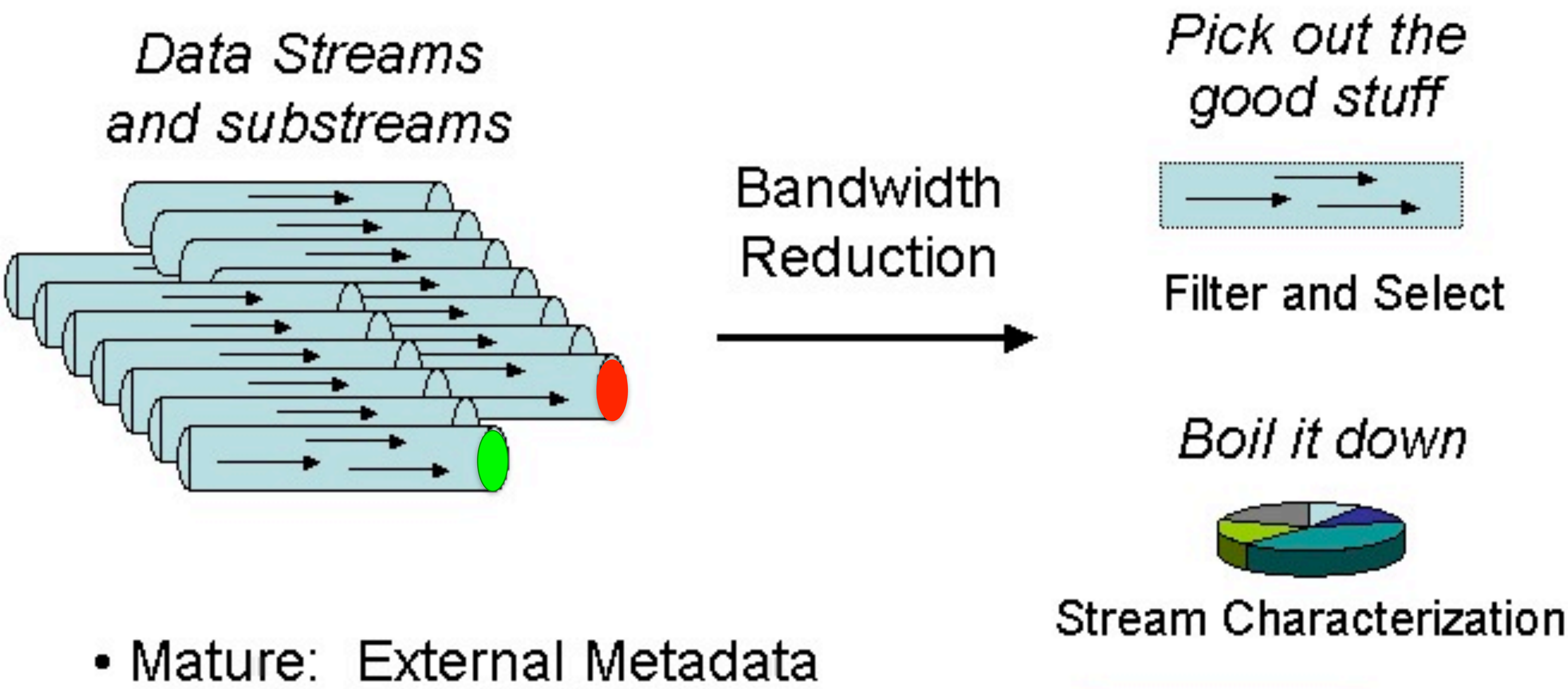
Coping with Information Overload



Coping with Information Overload

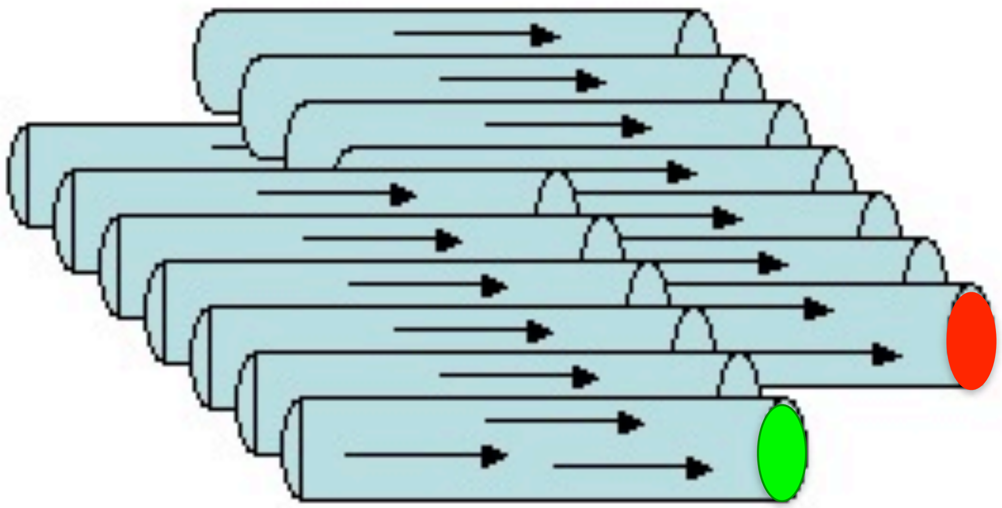


Coping with Information Overload



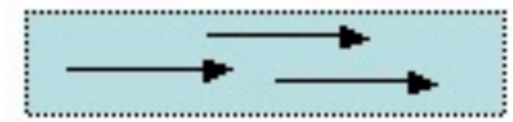
Coping with Information Overload

Data Streams and substreams



Bandwidth Reduction

Pick out the good stuff



Filter and Select

Boil it down

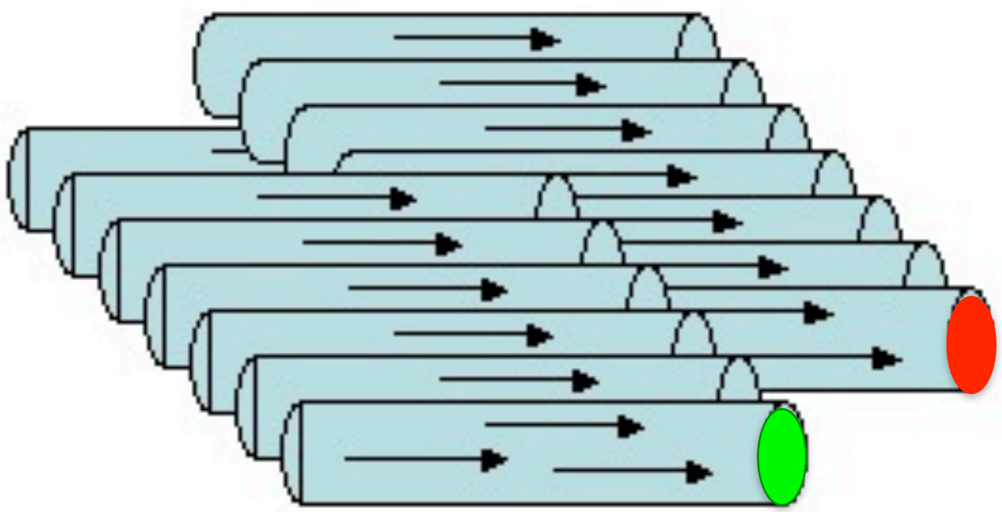
Stream Characterization

- Mature: External Metadata

- **Emerging: Metacontent**

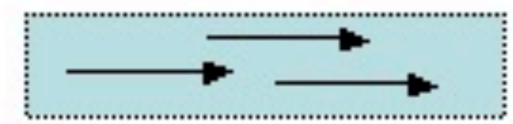
Coping with Information Overload

Data Streams and substreams



Bandwidth Reduction

Pick out the good stuff



Filter and Select

Boil it down

Stream Characterization

- Mature: External Metadata

- **Emerging: Metacontent**

- *language*
- *speaker*
- *topic*

Content has associated meta-data
challenge: how to exploit?

Content has associated meta-data
challenge: how to exploit?

- Associated meta-data is interpreted by humans as context

Content has associated meta-data *challenge: how to exploit?*

- Associated meta-data is interpreted by humans as context
- Humans acquire and use language during interaction with a complex environment, e.g. Roy's *Speechome* project at MIT

Content has associated meta-data *challenge: how to exploit?*

- Associated meta-data is interpreted by humans as context
- Humans acquire and use language during interaction with a complex environment, e.g. Roy's *Speechome* project at MIT
- Call centers have customer-profiles

Content has associated meta-data
challenge: how to exploit?

- Associated meta-data is interpreted by humans as context
- Humans acquire and use language during interaction with a complex environment, e.g. Roy's *Speechome* project at MIT
- Call centers have customer-profiles
- Voice messages have to/from telephone numbers

Content has associated meta-data
challenge: how to exploit?

- Associated meta-data is interpreted by humans as context
- Humans acquire and use language during interaction with a complex environment, e.g. Roy's *Speechome* project at MIT
- Call centers have customer-profiles
- Voice messages have to/from telephone numbers
- Enron email corpus has date, time, sender and recipients

Content has associated meta-data *challenge: how to exploit?*

- Associated meta-data is interpreted by humans as context
- Humans acquire and use language during interaction with a complex environment, e.g. Roy's *Speechome* project at MIT
- Call centers have customer-profiles
- Voice messages have to/from telephone numbers
- Enron email corpus has date, time, sender and recipients
- Switchboard dialog corpus has demographics: age, gender,….

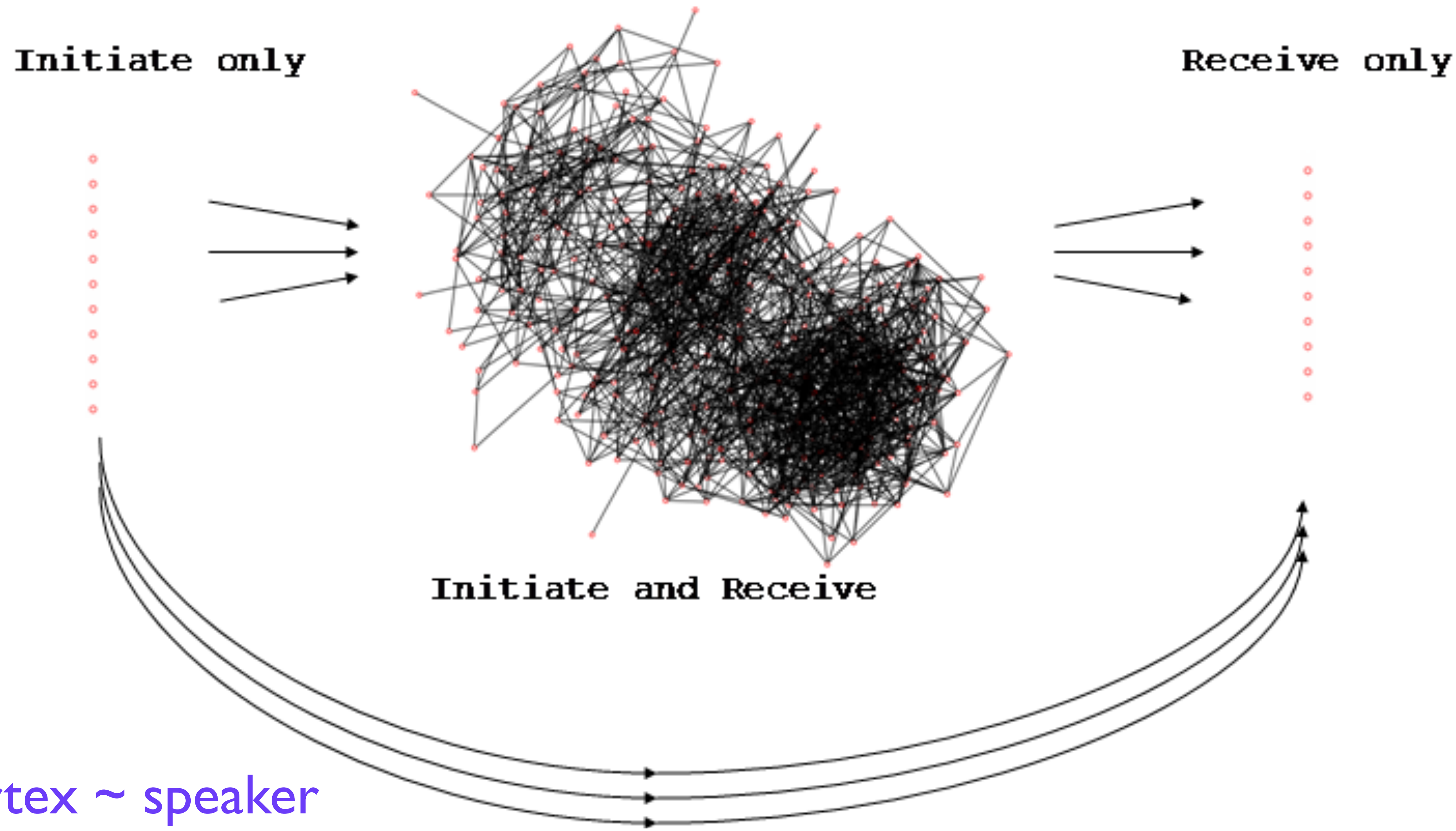
Content has associated meta-data *challenge: how to exploit?*

- Associated meta-data is interpreted by humans as context
- Humans acquire and use language during interaction with a complex environment, e.g. Roy's *Speechome* project at MIT
- Call centers have customer-profiles
- Voice messages have to/from telephone numbers
- Enron email corpus has date, time, sender and recipients
- Switchboard dialog corpus has demographics: age, gender,….
- Citeseer scientific articles have authors and citations

Communication Events from the Enron Corpus

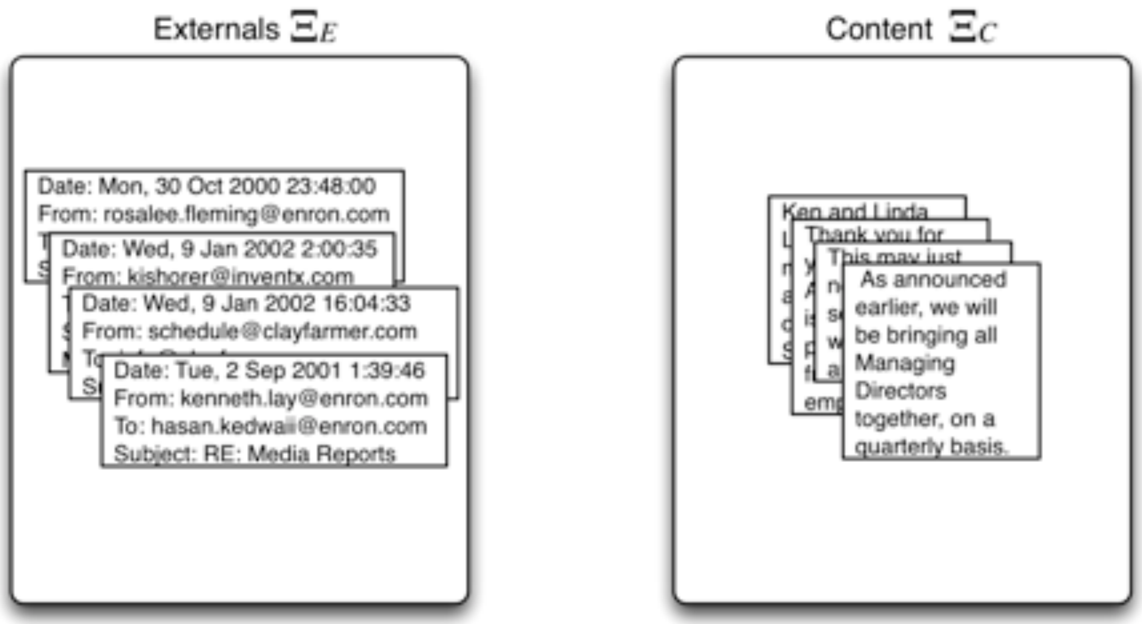
Date	Time	Sender	Receiver	Sender's Rank	Topic
2001-01-02	04:15:00	steven.k	jeff.d	Vice President	(1) California Analysis
2001-02-09	13:49:09	louise.k	andy.z	President	(9) Daily Business
2001-02-16	21:06:00	drew.f	jeff.d	Vice President	(5) California Enron
2001-02-26	22:30:00	james.s	john.l	Vice President	(14) Energy Newsfeed
2001-03-01	07:54:00	diana.s	kate.s	Trader	(5) California Enron
2001-04-06	05:15:00	mike.g	john.l	Manager	(7) Newsfeed California
2001-04-16	06:12:00	richard.s	steven.k	Vice President	(9) Daily Business
2001-05-11	16:02:00	andy.z	john.l	Vice President	(11) Enron Online
2001-06-27	17:44:24	s..s	geoff.s	Vice President	(9) Daily Business
2001-09-05	14:36:53	geoff.s	louise.k	Director	(12) Enrononline Daily
2001-09-15	20:51:20	m..p	louise.k	Vice President	(12) Enrononline Daily
2001-10-04	14:19:16	john.l	louise.k	CEO	(11) Enron Online
2001-10-05	18:49:05	j..k	richard.s	Vice President	(9) Daily Business
2001-10-08	17:50:19	shelley.c	darrell.s	Vice President	(1) California Analysis

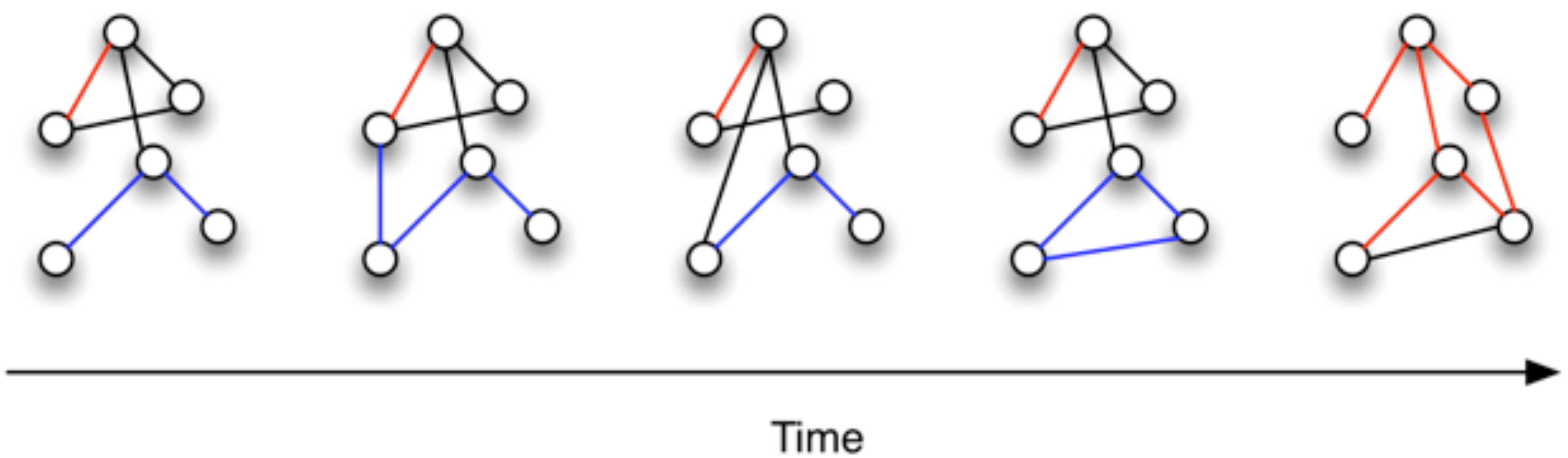
SwitchBoard Communications Graph



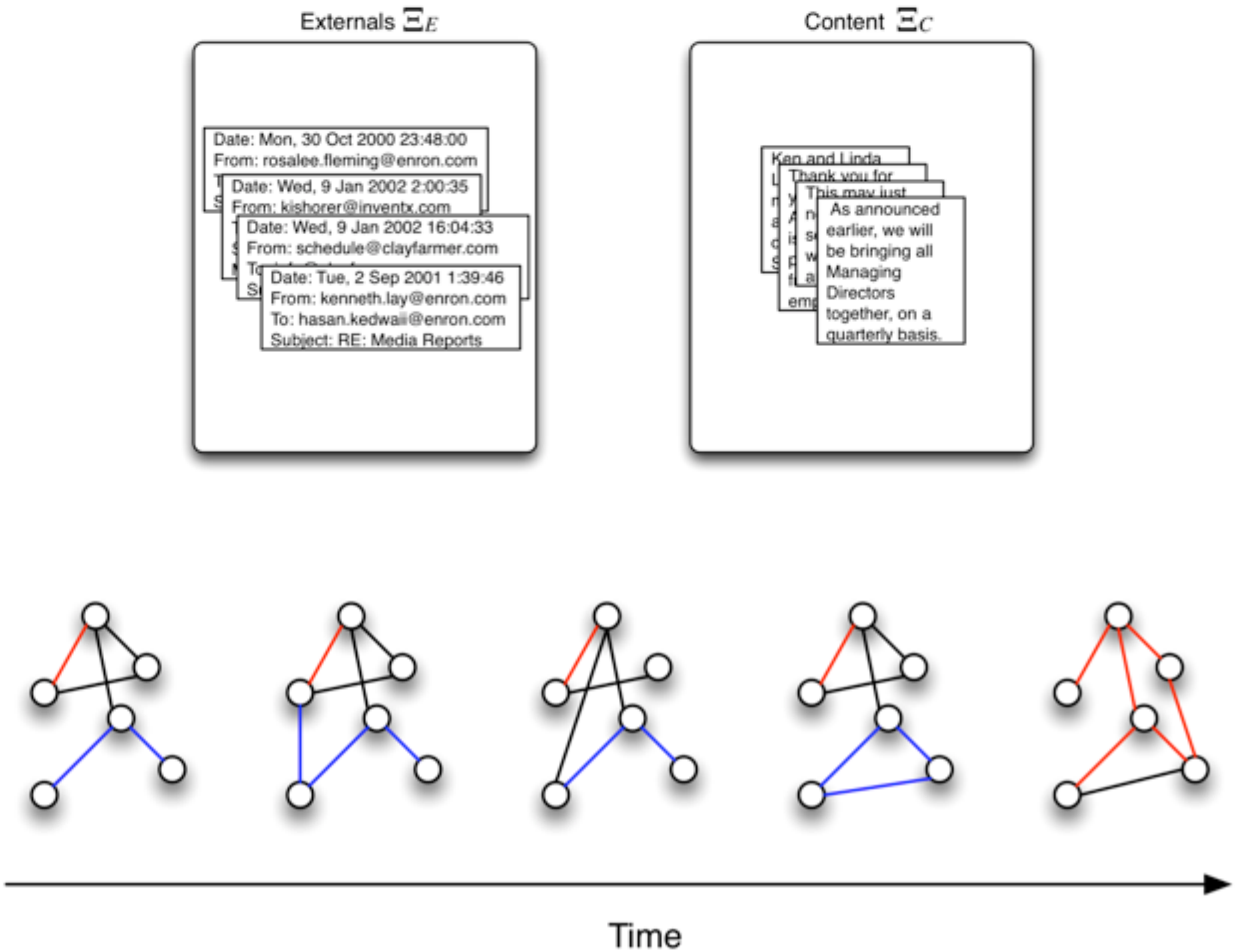
Vertex ~ speaker
Edge ~ dialog

Time Series of Attributed Graphs





Time Series of Attributed Graphs



Generated by some random process $\mathbf{G}_t$?

Random Attributed Graphs (RAGs)

Random Attributed Graphs (RAGs)

- There is significant literature on random graphs, ignoring conte***n***t .

Random Attributed Graphs (RAGs)

- There is significant literature on random graphs, ignoring conte***n***t .
- There is significant literature on stochastic models for language and documents streams, ignoring conte***x***t.

Random Attributed Graphs (RAGs)

- There is significant literature on random graphs, ignoring conte***n***t .
- There is significant literature on stochastic models for language and documents streams, ignoring conte***x***t.
- There is a computer science literature on attributed graphs, e.g. as produced by entity and relations, ignoring stochastic modeling.

Random Attributed Graphs (RAGs)

- There is significant literature on random graphs, ignoring conte***n***t .
- There is significant literature on stochastic models for language and documents streams, ignoring conte***x***t.
- There is a computer science literature on attributed graphs, e.g. as produced by entity and relations, ignoring stochastic modeling.
- Before this research effort, *no* literature that we know of addressing time series of random attributed graphs.

Generative Models for RAGs

- Build RAG models by extending random graph models
- Erdos-Renyi (binomial) graphs, where a pair of vertices is connected with *iid* probability *p*.
- Kidney/Egg models, Block models
- Latent Position and Random Dot Product Models where

$$p_{ij} = h(x_i , x_j)$$

- Construct from time series of communication events

$$M = \{ (t, u_t , v_t , s_t) \}_t$$

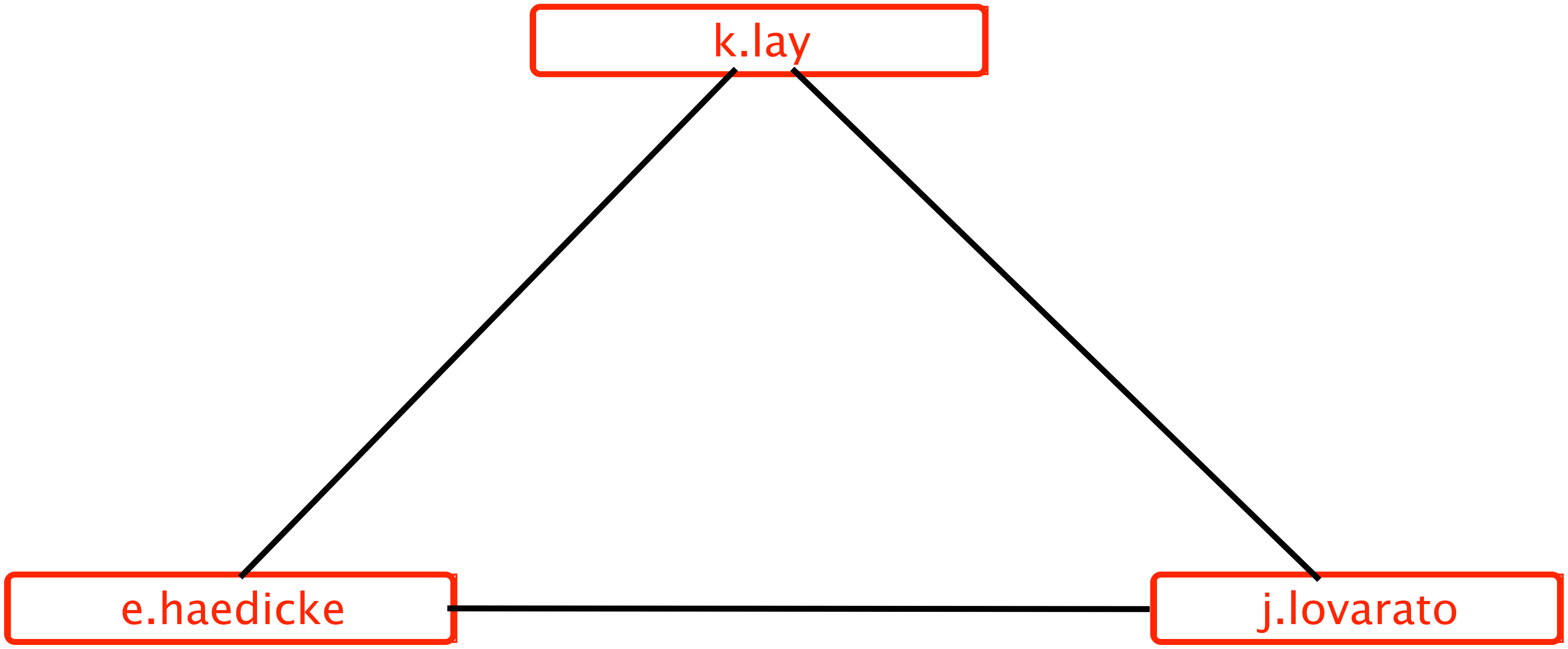
Vertex Nomination

- Cf. fraud and social network analysis
  - significant literature using graphs
- Intuition for fusion is clear
- Experimental evaluation on Enron email corpus
- Summer workshop
  - at JHU Human Language Technology COE
  - participants from all over the U.S.

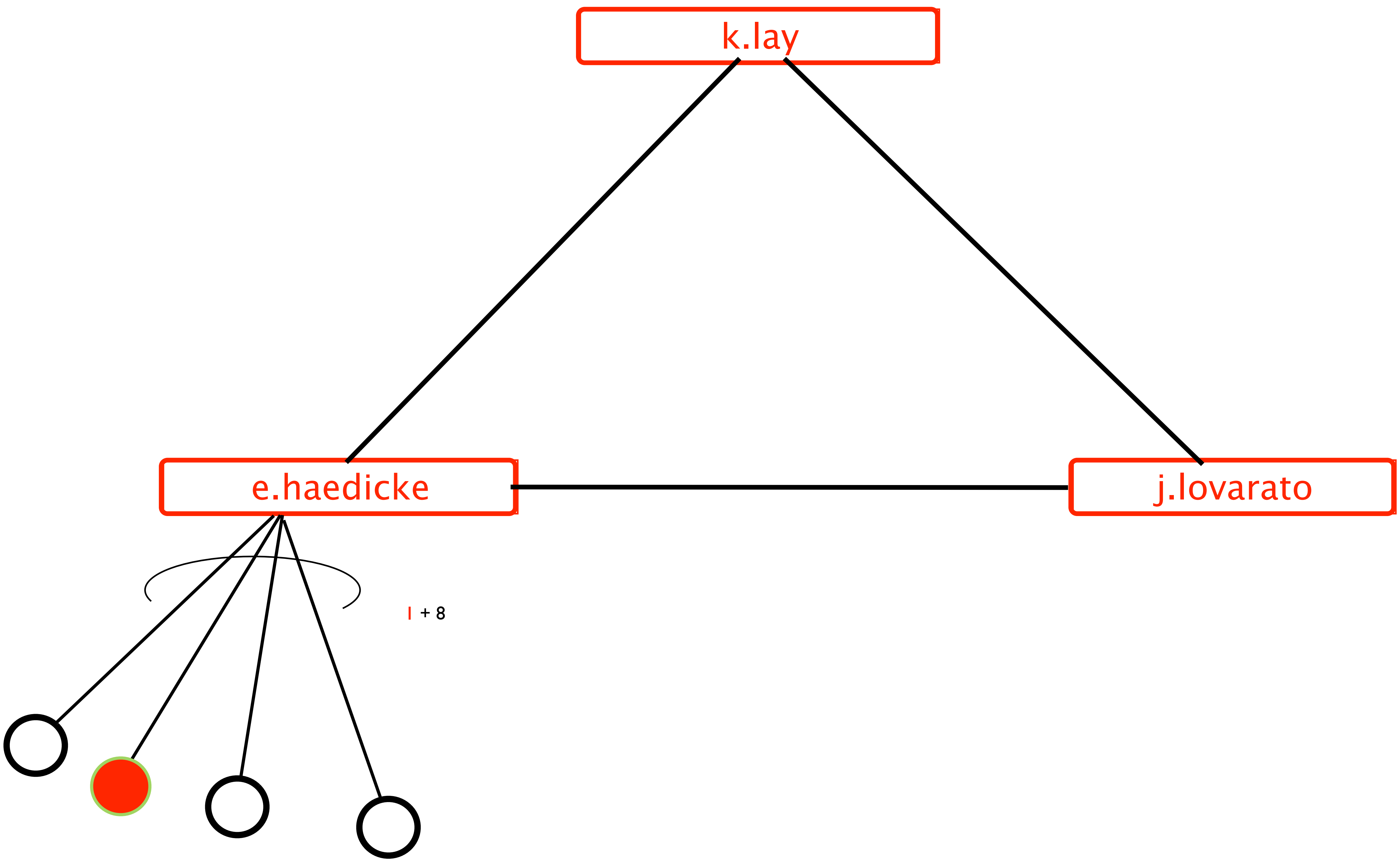
Experimental Methodology

- Given a set of red vertices
- Occlude subset of red vertices
- Develop method for nominating vertices as red
- Evaluate on how well it discovers those occluded *red* vertices
  - versus false nominations

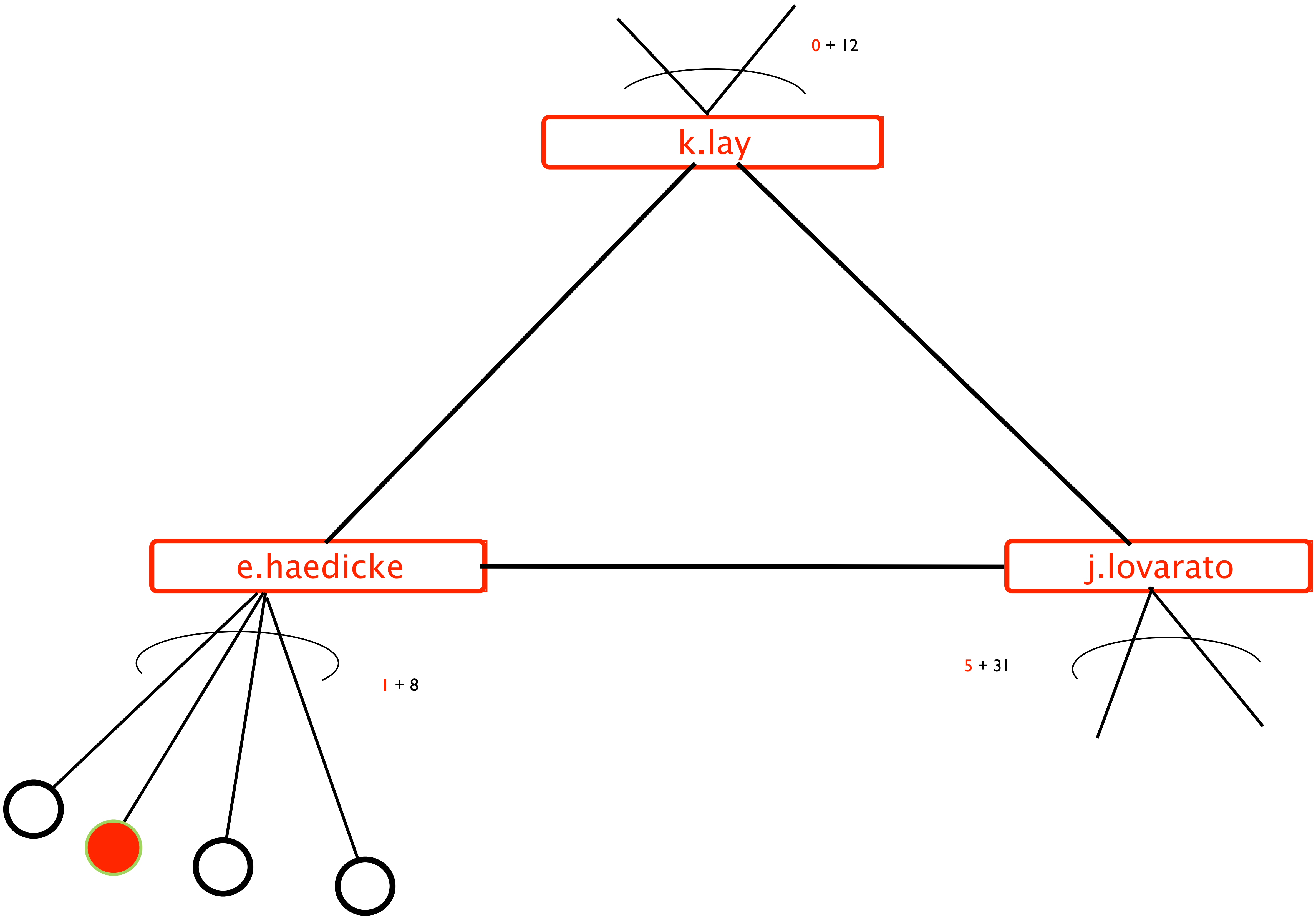
Enron Example: Red Vertices -> Red Documents



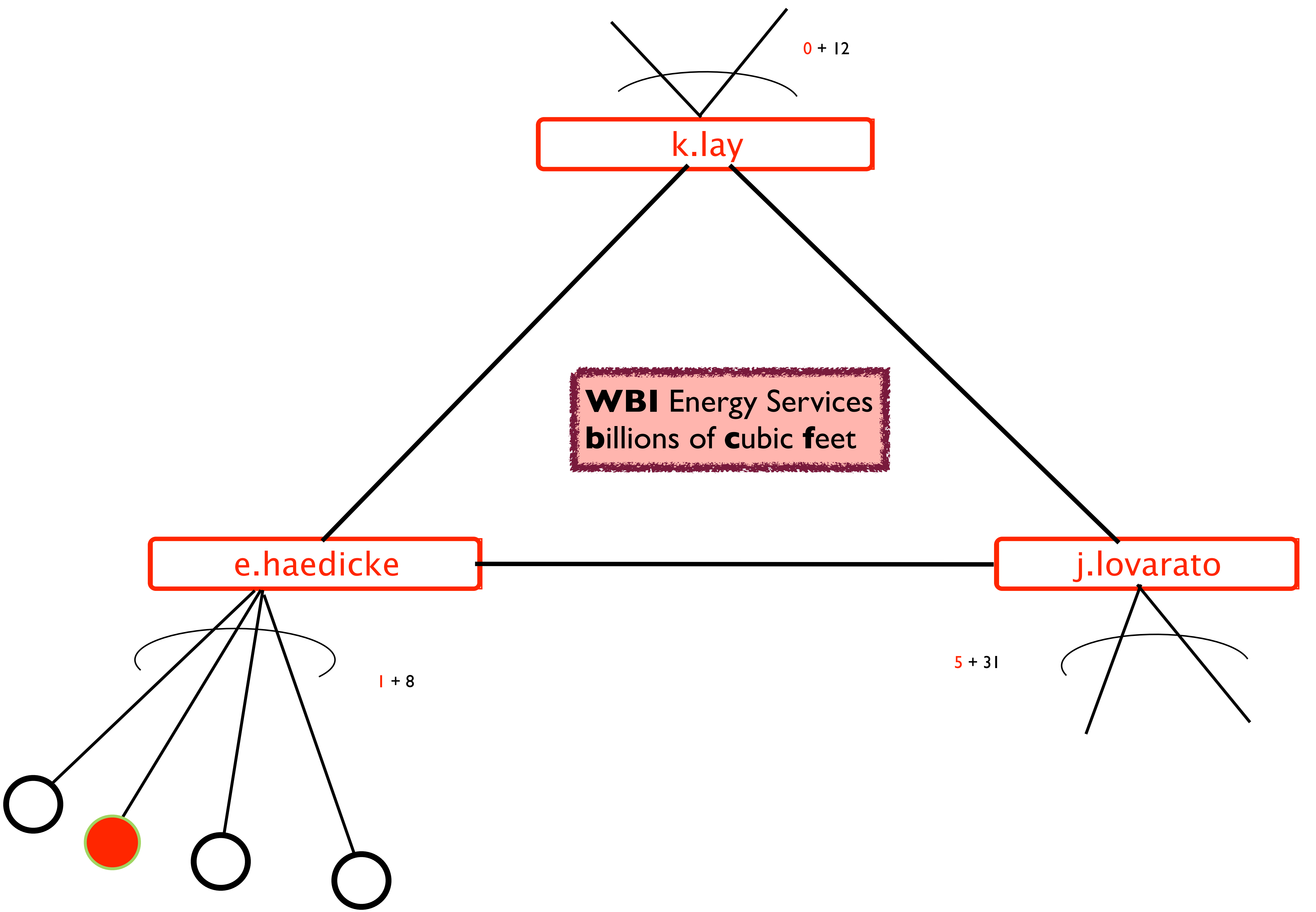
Enron Example: Red Vertices -> Red Documents



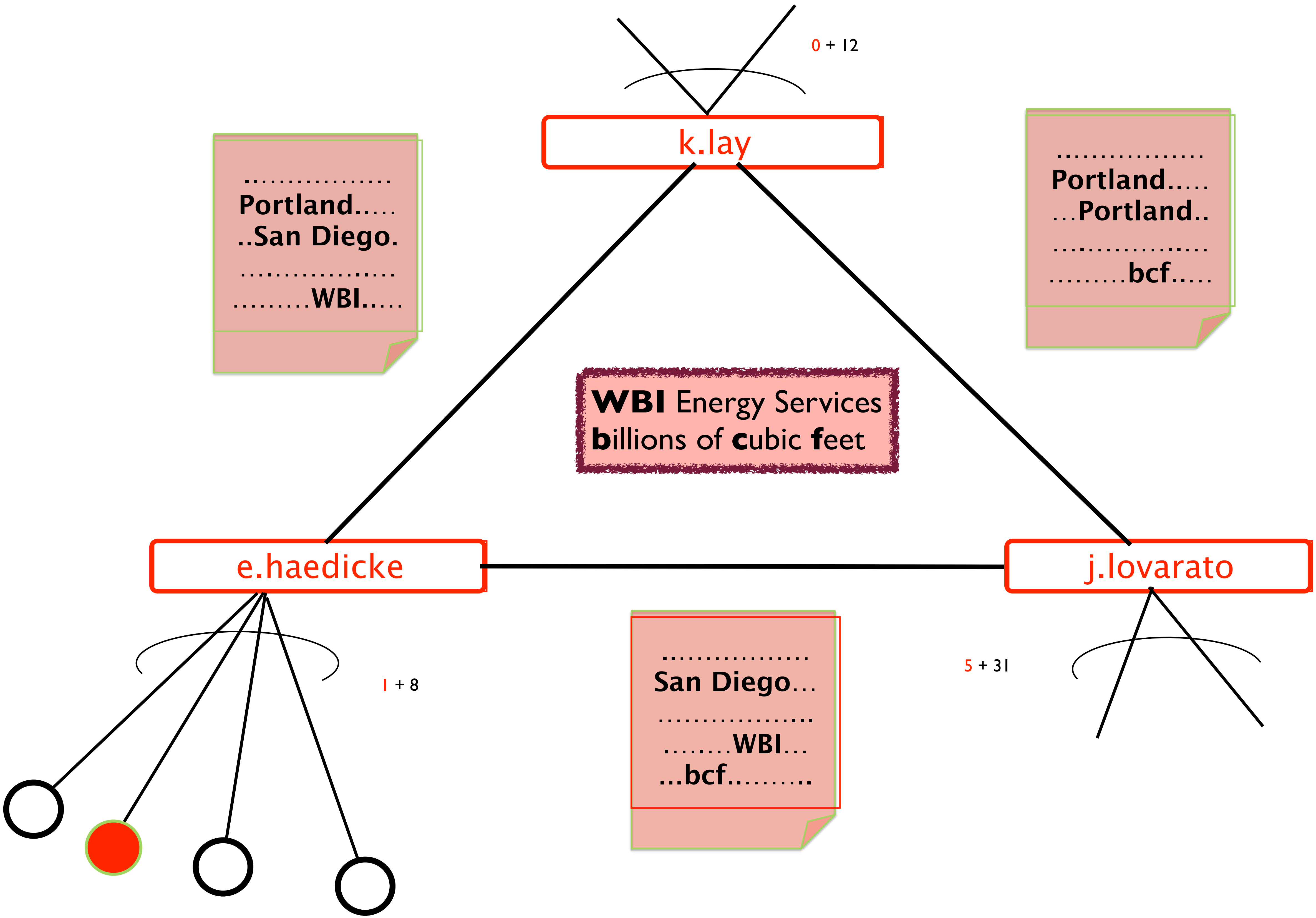
Enron Example: Red Vertices -> Red Documents



Enron Example: Red Vertices -> Red Documents



Enron Example: Red Vertices -> Red Documents



Edge Attributed Graph -> Latent Vertex Attributes

Edge Attributed Graph -> Latent Vertex Attributes

- From red vertices, now have induced 'red topic model'

Edge Attributed Graph -> Latent Vertex Attributes

- From red vertices, now have induced 'red topic model'
- Use red model which is used to attribute all edges in the graph

Edge Attributed Graph -> Latent Vertex Attributes

- From red vertices, now have induced 'red topic model'
- Use red model which is used to attribute all edges in the graph
- Estimate latent vertex attributes for an attributed rdp model that best fit this attributed graph

Edge Attributed Graph -> Latent Vertex Attributes

- From red vertices, now have induced ‘red topic model’
- Use red model which is used to attribute all edges in the graph
- Estimate latent vertex attributes for an attributed rdp model that best fit this attributed graph
- The vertex attributes are x_0, x_1, x_2, where

Edge Attributed Graph -> Latent Vertex Attributes

- From red vertices, now have induced 'red topic model'
- Use red model which is used to attribute all edges in the graph
- Estimate latent vertex attributes for an attributed rdp model that best fit this attributed graph
- The vertex attributes are x_0, x_1, x_2, where
  - $\mathbf{x_1}$ is the tendency of the vertex to engage in red communications

Edge Attributed Graph -> Latent Vertex Attributes

- From red vertices, now have induced 'red topic model'
- Use red model which is used to attribute all edges in the graph
- Estimate latent vertex attributes for an attributed rdp model that best fit this attributed graph
- The vertex attributes are x_0, x_1, x_2, where
  - $\mathbf{x_1}$ is the tendency of the vertex to engage in red communications
    - abuse jargon and call this the 'redness of the vertex'

Edge Attributed Graph -> Latent Vertex Attributes

- From red vertices, now have induced 'red topic model'
- Use red model which is used to attribute all edges in the graph
- Estimate latent vertex attributes for an attributed rdp model that best fit this attributed graph
- The vertex attributes are x_0, x_1, x_2, where
  - $\mathbf{x_1}$ is the tendency of the vertex to engage in red communications
    - abuse jargon and call this the 'redness of the vertex'
  - $\mathbf{x_2}$ is the tendency to engage in non-red communication

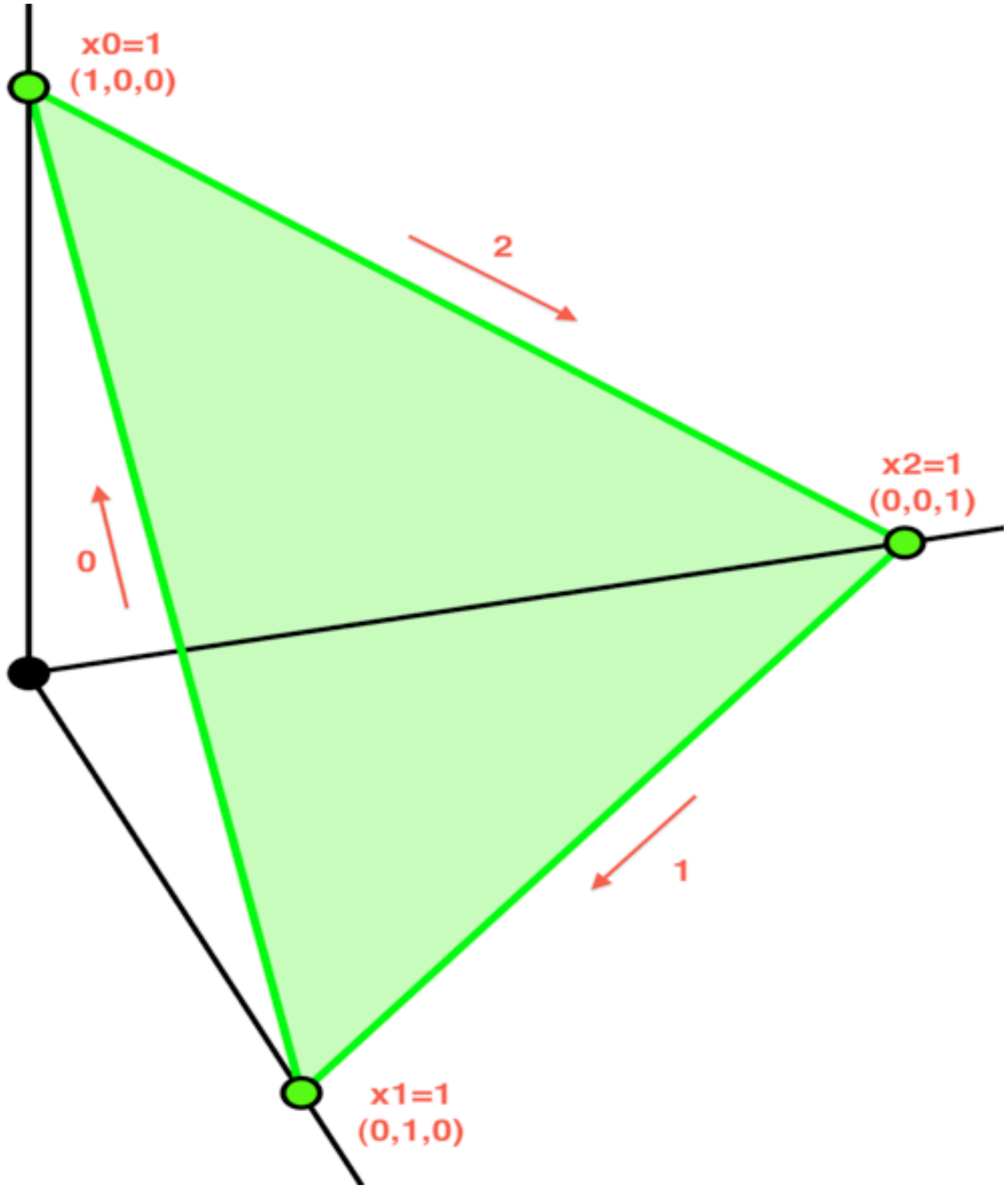
Edge Attributed Graph -> Latent Vertex Attributes

- From red vertices, now have induced ‘red topic model’
- Use red model which is used to attribute all edges in the graph
- Estimate latent vertex attributes for an attributed rdp model that best fit this attributed graph
- The vertex attributes are x_0, x_1, x_2, where
  - $\mathbf{x_1}$ is the tendency of the vertex to engage in red communications
    - abuse jargon and call this the ‘redness of the vertex’
  - $\mathbf{x_2}$ is the tendency to engage in non-red communication
    - call this the ‘greenness of the vertex’

Edge Attributed Graph -> Latent Vertex Attributes

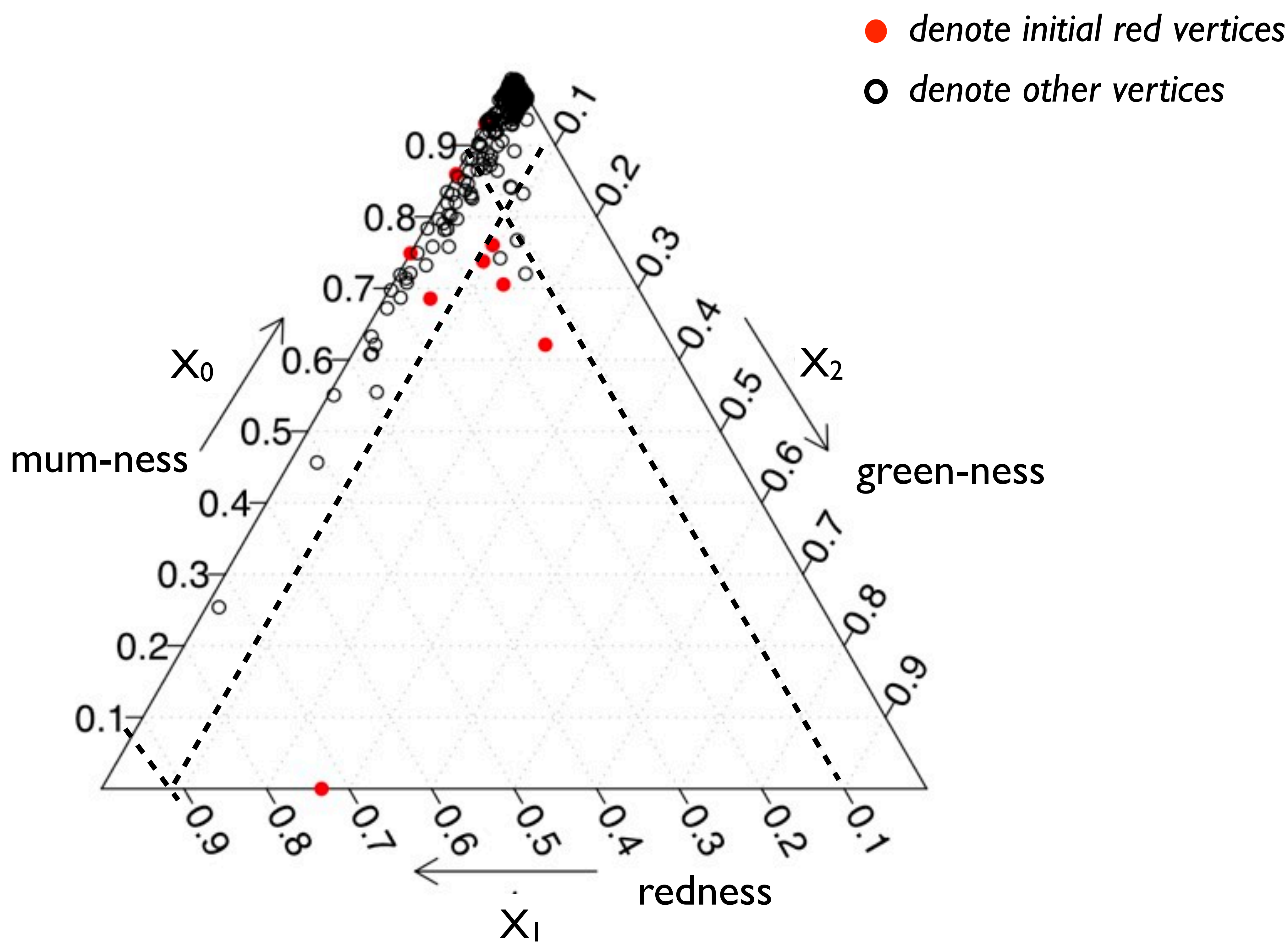
- From red vertices, now have induced 'red topic model'
- Use red model which is used to attribute all edges in the graph
- Estimate latent vertex attributes for an attributed rdp model that best fit this attributed graph
- The vertex attributes are x_0, x_1, x_2, where
  - $\mathbf{x_1}$ is the tendency of the vertex to engage in red communications
    - abuse jargon and call this the 'redness of the vertex'
  - $\mathbf{x_2}$ is the tendency to engage in non-red communication
    - call this the 'greenness of the vertex'
  - $\mathbf{x_0} = 1 - x_1 - x_2$ = non-edginess = tendency of the vertex to stay mum

Latent Vertex Attributes live in the 2D simplex



x0 ~ mum-ness
x1 ~ redness
x2 ~ greenness

Distribution of 184 Latent Vertex Attributes



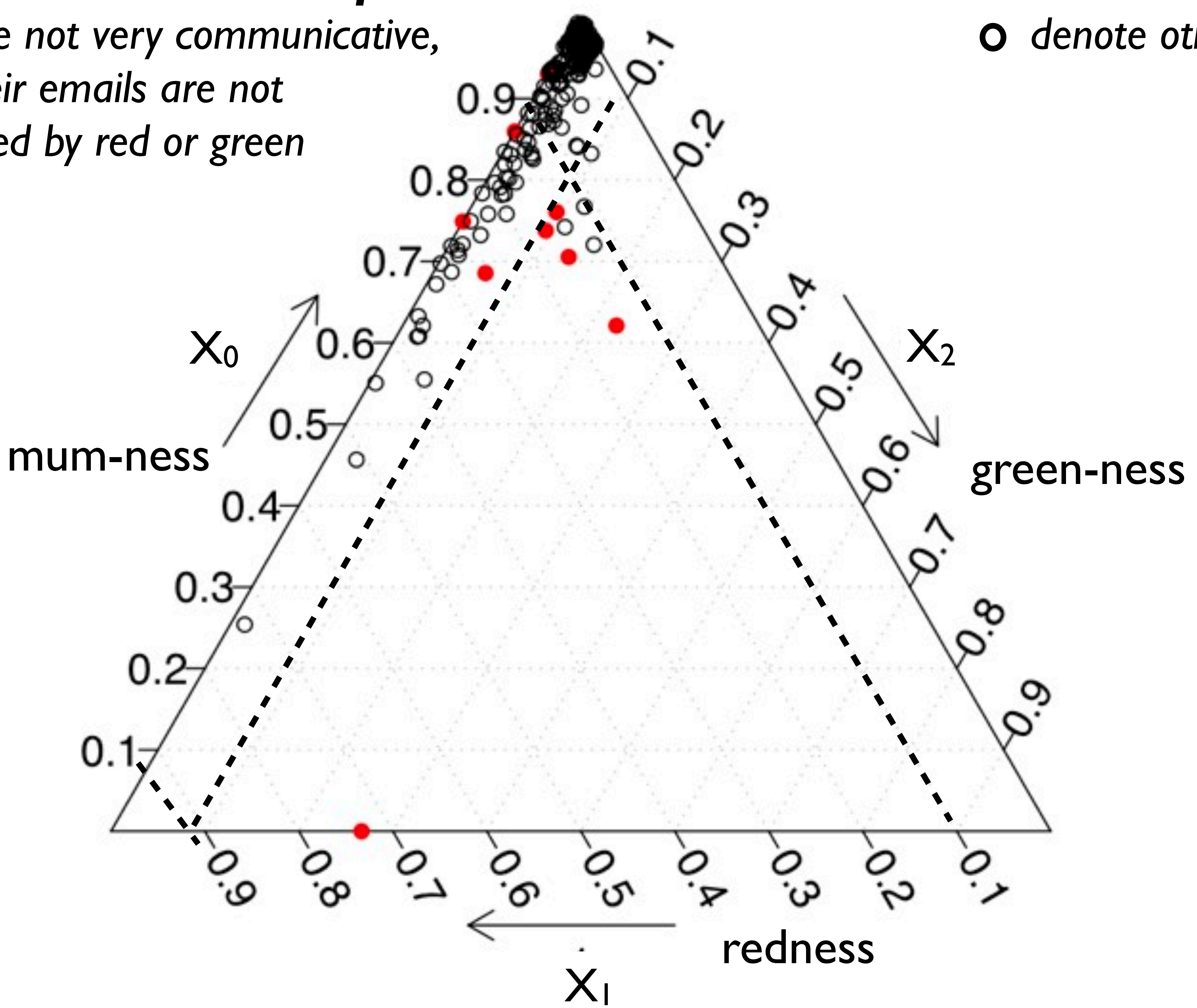
Distribution of 184 Latent Vertex Attributes

Sparse Communication Graph

most vertices are not very communicative, and their emails are not dominated by red or green

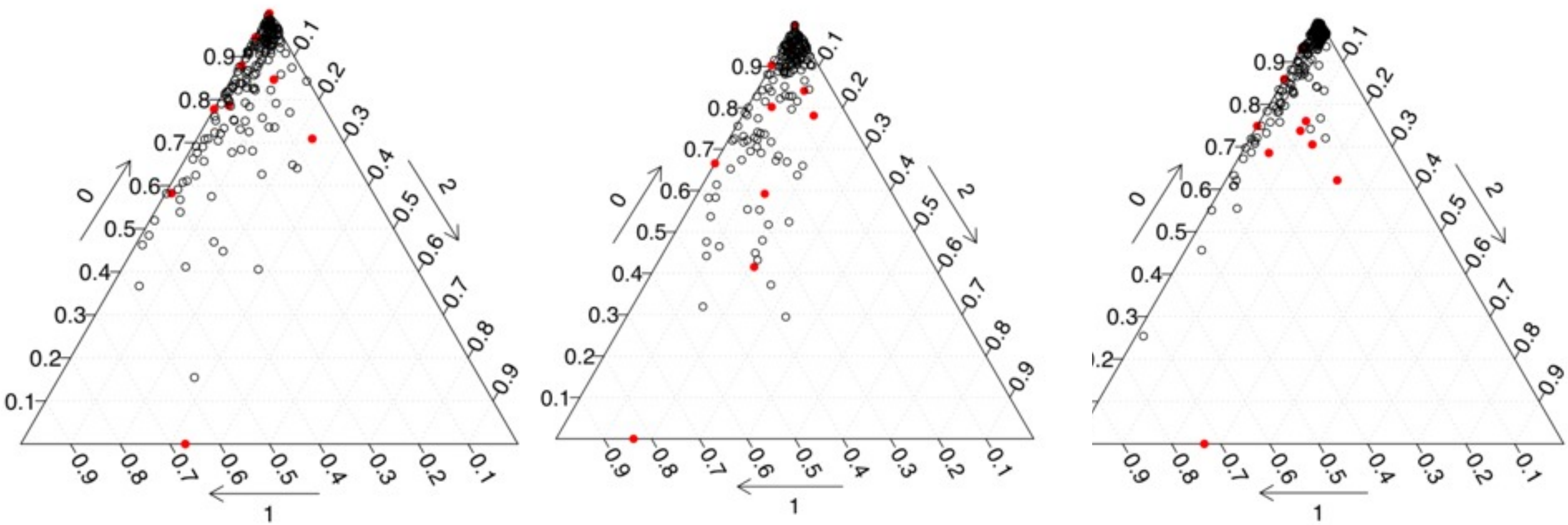
denote initial red vertices

denote other vertices



Anomalous Chatter Group in Enron Time Series

Induced Egg



Egg?	p>0.99	p~ 0.7	**p < 0.01**
Time	Weeks 18-37	Weeks 38-57	Weeks 58-77

Conclusions

- New Methods for Fusion of Conte***x***t and Conte***n***t
- Pioneered at JHU Human Language Technology COE
- Theory, Algorithms and Experimental Evaluation
- Tasks
  - Stream Characterization
  - Vertex Nomination
  - Dyadic Priors
- Experimentally evaluated on
  - Enron email corpus
  - Switchboard speech corpus
  - other data

Some References

- ***Statistical Inference on Random Graphs: Fusion of Graph Features and Content,*** Grothendieck, Priebe, and Gorin, Computational Statistics and Data Analysis (2010)
- ***Statistical Inference on random attributed Graphs: Fusion of Graph Features and Content: An Experiment on Time-series of Enron Graphs,*** Priebe et al, Computational Statistics and Data Analysis (2010).
- ***Towards Link Characterization from Content: Recovering Distributions from Classifier Output***, Grothendieck and Gorin , IEEE Transactions on Speech and Audio, May 2008
- ***Vertex Nomination via Content and Context,*** *Coppersmith and Priebe submitted for publication*
- ***Vertex Nomination via Attributed Random Dot Product Graphs****, Marchette, Priebe, Coppersmith , Proc. International Statistical Institute, 2011.*
- ***Latent Process Model for Time Series of Attributed Random Graphs****, Lee and Priebe, Statistical Inference for Stochastic Processes, 2011*