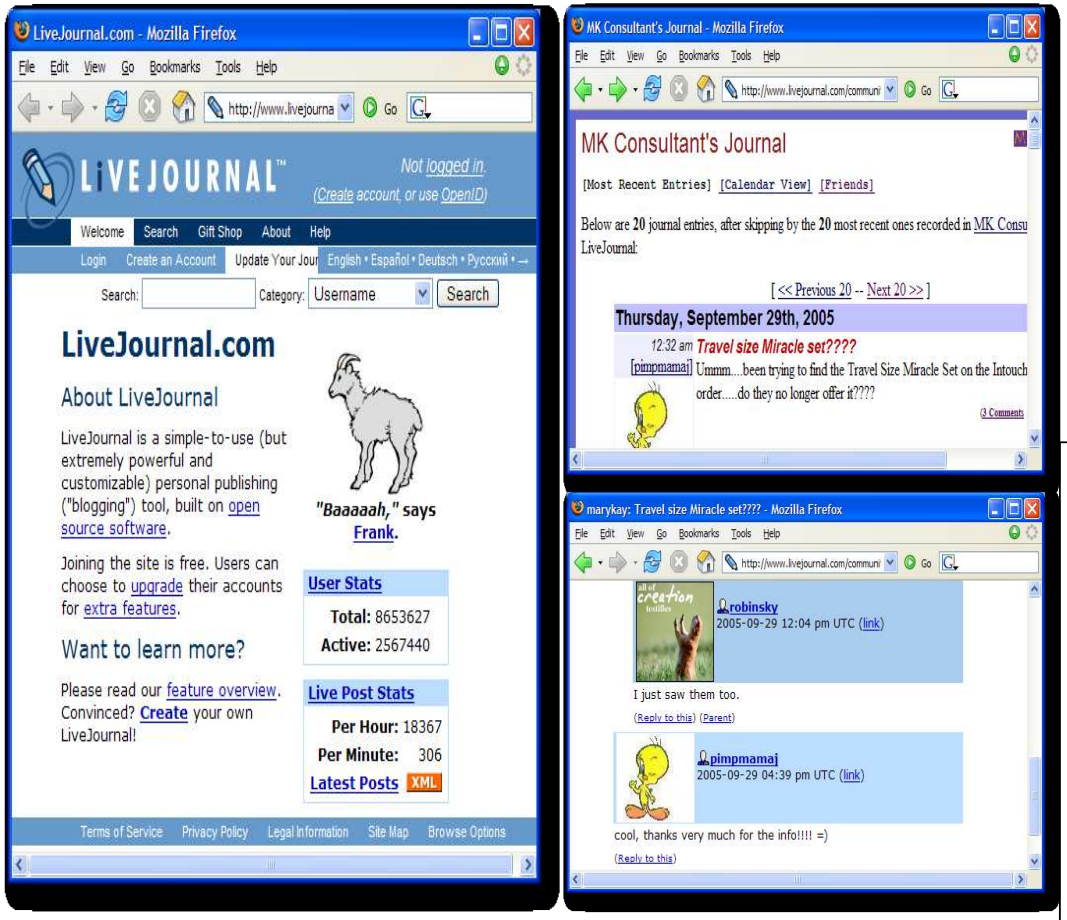
Finding Hidden Group Structure in a Stream of Communications

J. Baumes, M. Goldberg, M. Hayvanovych, M. Magdon-Ismail,
W. Wallace, M. Zaki

May 23, 2006.

Communication Streams



Time: January 12, 2005, 09:35

From: joe@xyz.com

To: sue@abc.com

Subject: Hello

Message: Where have you been?

16:06:31] <FreeTrade> Republicans were the worst pacifists before ww1 and ww2
[16:06:43] <SweetLeaf> France Fries
[16:06:50] <FreeTrade> As a generality, of course their were Republican Hawks.
[16:07:13] <FreeTrade> Sweet, good pun but bad story!
[16:07:18] <SweetLeaf> yup
[16:07:23] <Lupine> anyways, he's perpetually tormented by presidential actions
[16:07:25] <SweetLeaf> it aint good for no one
[16:07:47] <SweetLeaf> I think they knew it was commiing
[16:07:51] <FreeTrade> Rossevelt met monthly in New York with mostly trusted Republicans to talk about how to get america into the war.
[16:08:10] <FreeTrade> and he spent 2 year with Churchill meeting him sometimes secretly in the ocean to discuss the same topic.
[16:08:22] <FreeTrade> Exchanging a lot of letters.
[16:08:25] <FreeTrade> telegrams
[16:08:28] <Lupine> There really is nothing like a shorn scrotum. It's breathtaking, I suggest you try it.
[16:08:55] <FreeTrade> Well they didnt literally meet in the ocean, they were on ships.

A Streaming Hidden Group

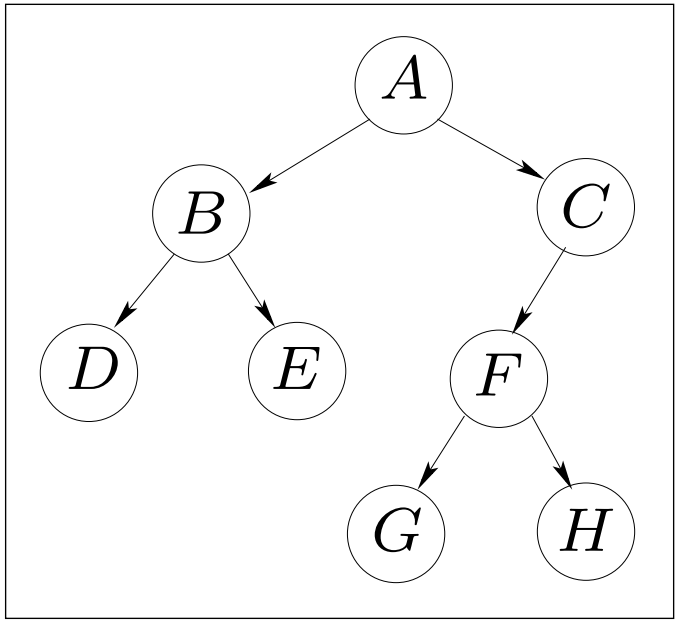
00	**A→C**	Golf tomorrow? Tell everyone.
05	**C→F**	Alice mentioned golf tomorrow.
06	**A→B**	Hey, golf tomorrow? Spread the word
12	**A→B**	Tee time: 8am; Place: Pinehurst.
13	**F→G**	Hey guys, golf tomorrow .
13	**F→H**	Hey guys, golf tomorrow .
15	**A→C**	Tee time: 8am; Place: Pinehurst.
20	**B→D**	We're playing golf tomorrow.
20	**B→E**	We're playing golf tomorrow.
22	**C→F**	Tee time: 8am; Place: Pinehurst.
25	**B→D**	Tee time: 8am; Place: Pinehurst.
25	**B→E**	Tee time 8am, Pinehurst.
31	**F→G**	Tee time 8am, Pinehurst.
31	**F→H**	Tee off 8am,Pinehurst.

Two Communication Waves

00	**A→C**	Golf tomorrow? Tell everyone.
05	**C→F**	Alice mentioned golf tomorrow.
06	**A→B**	Hey, golf tomorrow? Spread the word
12	**A→B**	Tee time: 8am; Place: Pinehurst.
13	**F→G**	Hey guys, golf tomorrow .
13	**F→H**	Hey guys, golf tomorrow .
15	**A→C**	Tee time: 8am; Place: Pinehurst.
20	**B→D**	We're playing golf tomorrow.
20	**B→E**	We're playing golf tomorrow.
22	**C→F**	Tee time: 8am; Place: Pinehurst.
25	**B→D**	Tee time: 8am; Place: Pinehurst.
25	**B→E**	Tee time 8am, Pinehurst.
31	**F→G**	Tee time 8am, Pinehurst.
31	**F→H**	Tee off 8am,Pinehurst.

Inferred Group Structure

00	**A→C**	Golf tomorrow? Tell everyone.
05	**C→F**	Alice mentioned golf tomorrow.
06	**A→B**	Hey, golf tomorrow? Spread the word
12	**A→B**	Tee time: 8am; Place: Pinehurst.
13	**F→G**	Hey guys, golf tomorrow .
13	**F→H**	Hey guys, golf tomorrow .
15	**A→C**	Tee time: 8am; Place: Pinehurst.
20	**B→D**	We're playing golf tomorrow.
20	**B→E**	We're playing golf tomorrow.
22	**C→F**	Tee time: 8am; Place: Pinehurst.
25	**B→D**	Tee time: 8am; Place: Pinehurst.
25	**B→E**	Tee time 8am, Pinehurst.
31	**F→G**	Tee time 8am, Pinehurst.
31	**F→H**	Tee off 8am,Pinehurst.



Challenge: No Semantic Information

00 **A→C**
05 **C→F**
06 **A→B**
12 **A→B**
13 **F→G**
13 **F→H**
15 **A→C**
20 **B→D**
20 **B→E**
22 **C→F**
25 **B→D**
25 **B→E**
31 **F→G**
31 **F→H**

Why: semantic information is hard to process; different languages; encrypted contents;...

Groups must communicate!

.

Background

1. **[MGWS03, BGMW04]** Finds such groups when waves do not overlap (cycle model).

2. **[BGM05, CSCC04, N03, GN02, BGKMP05, BGHMWZ06]** Other realted work on finding Hidden Structures (static groups/clustering).

Streaming Hidden Groups

1. Communication waves may **overlap** - bursty communication.

2. Waves may have different durations - propagation delays may be large or small.

3. Waves may propagate though the tree in different orders.

Problem Statement

00	**A**→**C**
05	**C**→**F**
06	**A**→**B**
12	**A**→**B**
13	**F**→**G**
13	**F**→**H**
15	**A**→**C**
20	**B**→**D**
20	**B**→**E**
22	**C**→**F**
25	**B**→**D**
25	**B**→**E**
31	**F**→**G**
31	**F**→**H**

Real Data

ENRON

Time: January 12, 2005, 09:35

From: joe@xyz.com

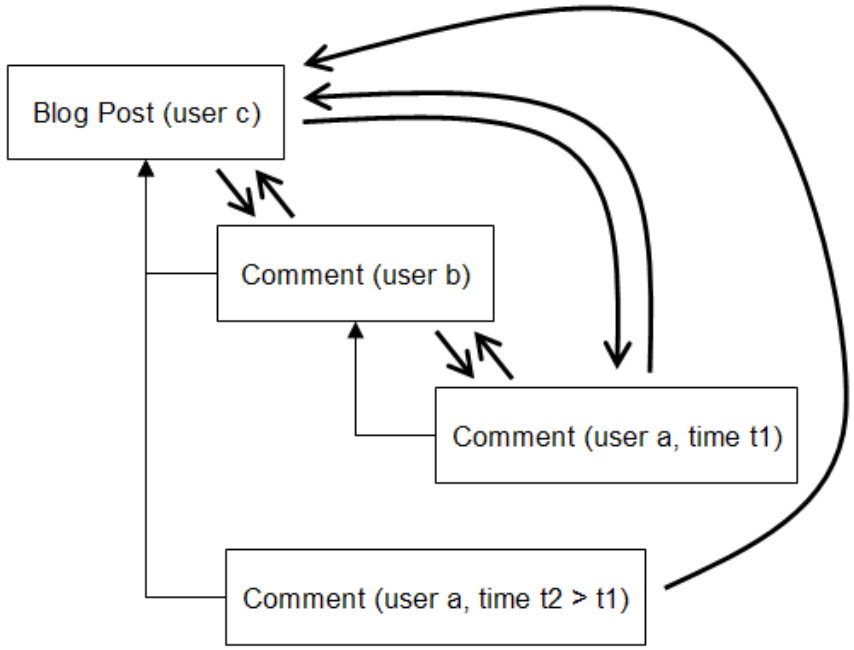
To: sue@abc.com

Subject: Hello

Message: Where have you been?

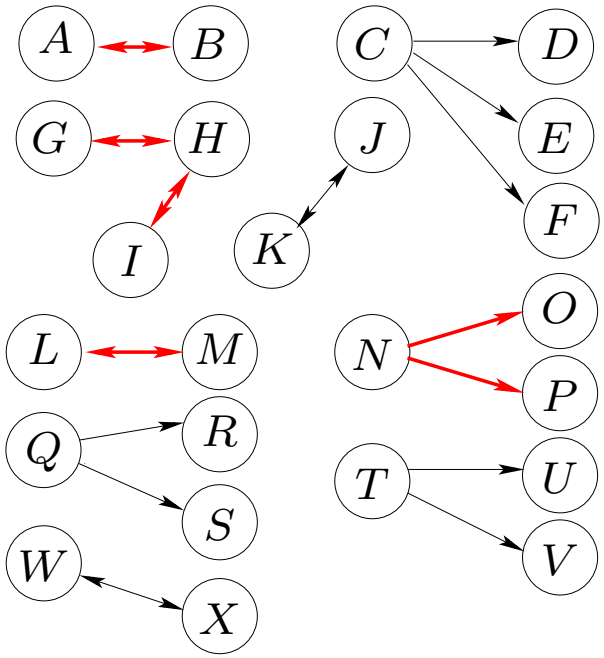
<January 12 2005 09:35; Joe; Sue>

WebLogs (Blogs)

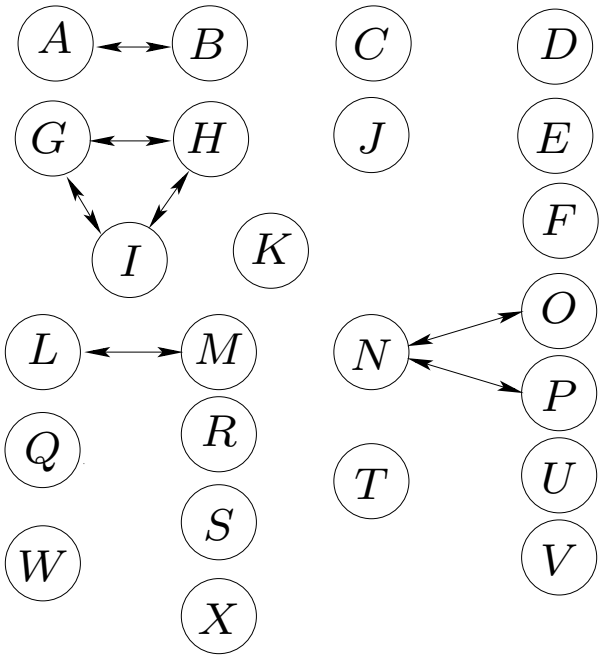


$< t_1, a, b >$
$< t_1, b, a >$
$< t_1, a, c >$
$< t_1, c, a >$
$< t_2, a, c >$

Weblog Results (Feb. 2006)



Discovered Weblog Groups

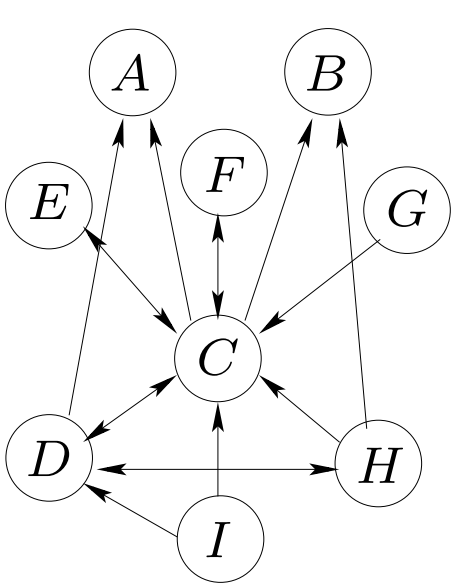


Published Friendship Graph

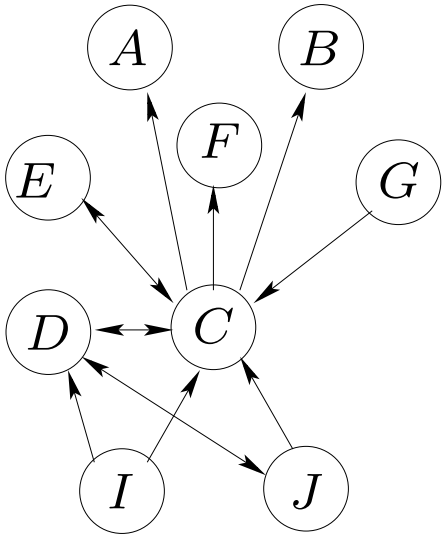
Probability of a random edge $P_{edge} = 0.0008\%$

Found Group = 2.5%

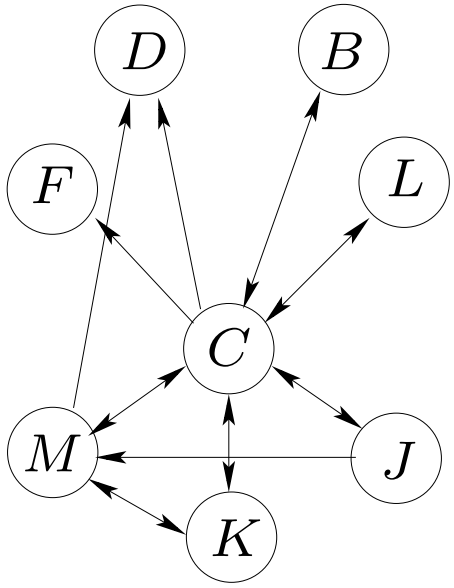
ENRON Evolution



Sept.2000 - Sept.2001



Mar.2001 - Mar.2002

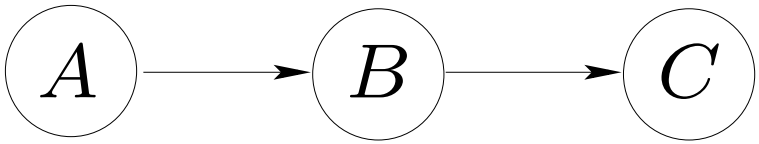


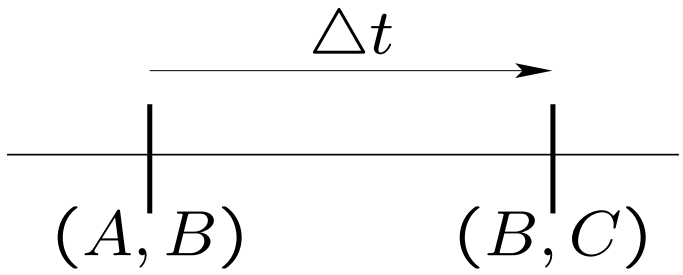
Sept.2001 - Sept.2002

Reminder: Problem Statement

00	**A**→**C**
05	**C**→**F**
06	**A**→**B**
12	**A**→**B**
13	**F**→**G**
13	**F**→**H**
15	**A**→**C**
20	**B**→**D**
20	**B**→**E**
22	**C**→**F**
25	**B**→**D**
25	**B**→**E**
31	**F**→**G**
31	**F**→**H**

Communication Propagation

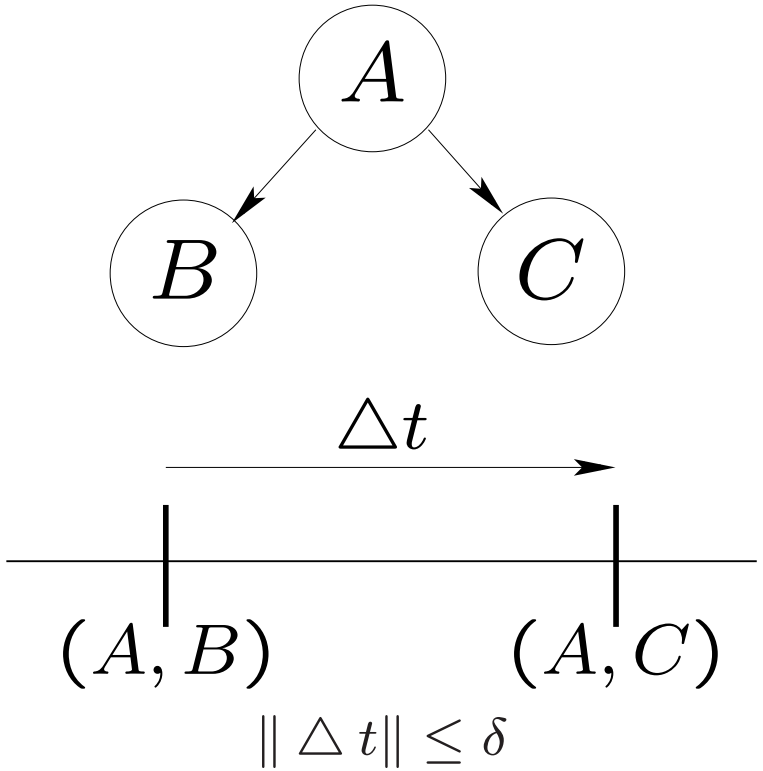




$$\tau_{min} \leq \triangle t \leq \tau_{max}$$

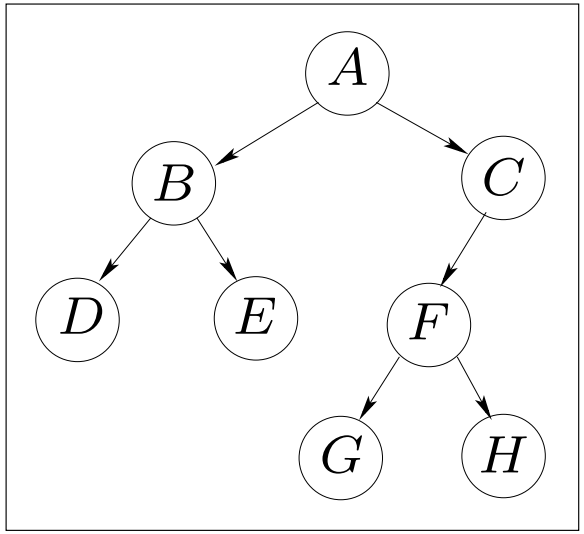
Chain triple

Communication Divergence



Sibling triple

Tree Occurance

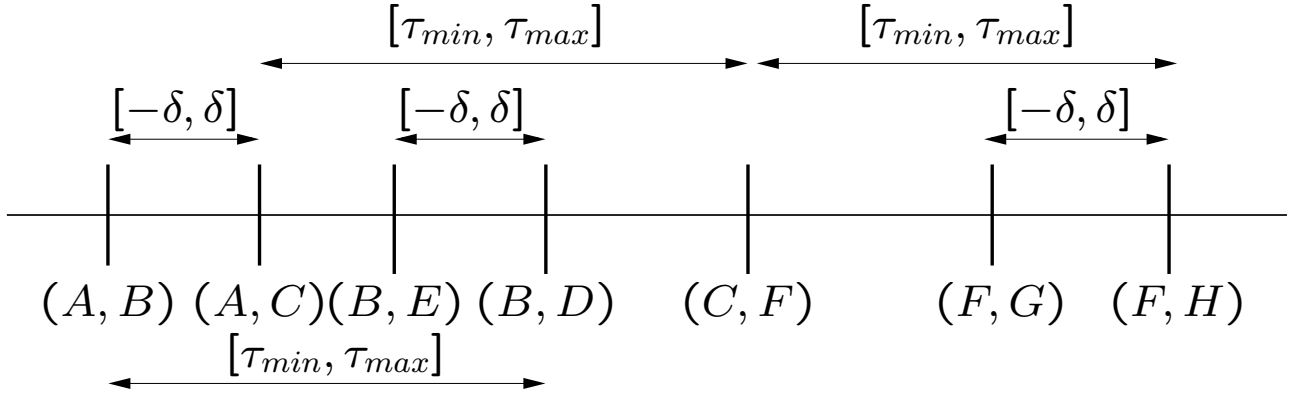


Chains:

$A \to B \to D$
$A \to B \to E$
$A \to C \to F$
$C \to F \to G$
$C \to F \to H$

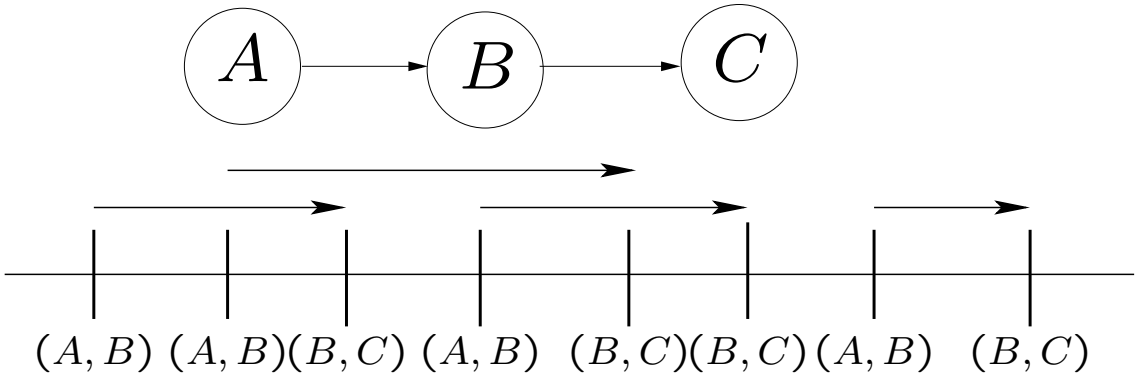
Siblings:

$A \to (B, C)$
$B \to (D, E)$
$F \to (G, H)$

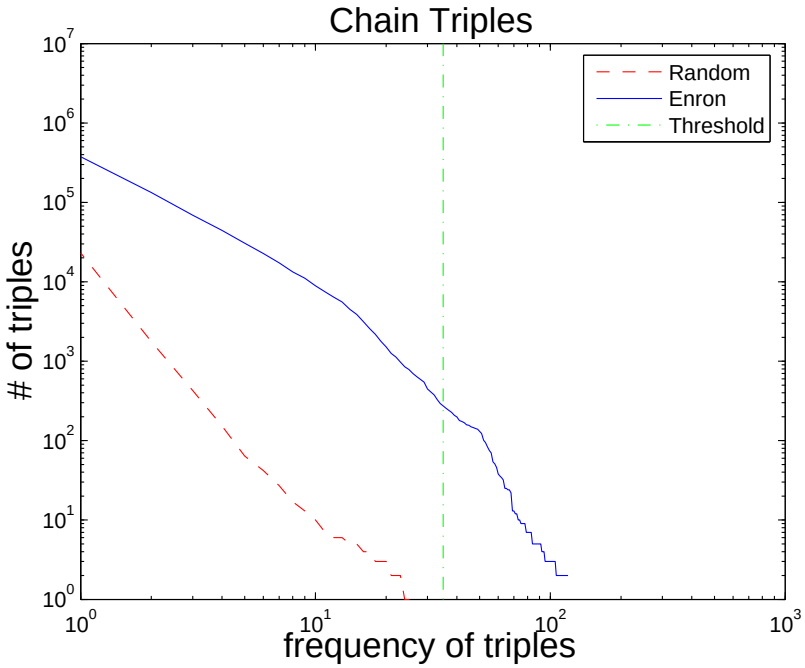


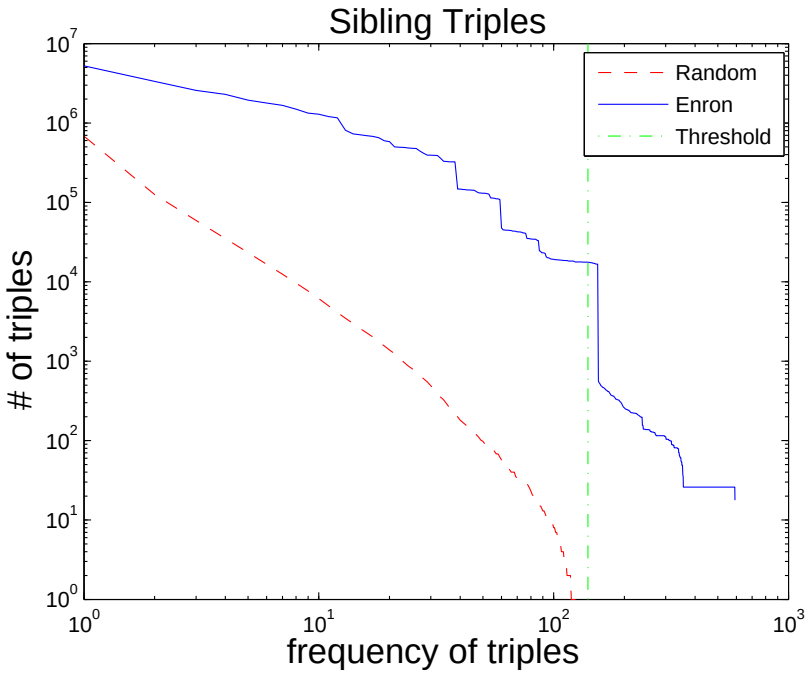
Tree occurs if every triple occurs

Suspicious Triples

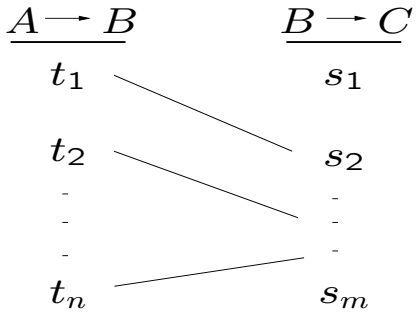


Triple which appears many times





Finding Suspicious Triples



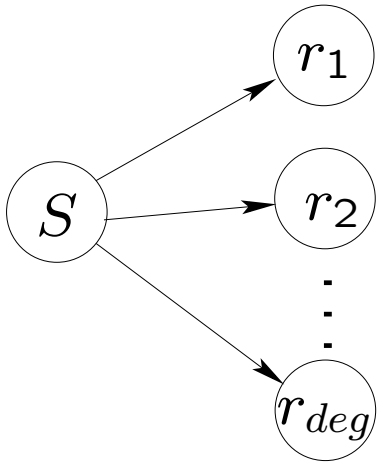
1. Find max. # occurrances **(2D max. matching)**.

2. **Suspicious** if match size $> \kappa_{sig}$.

Algorithms

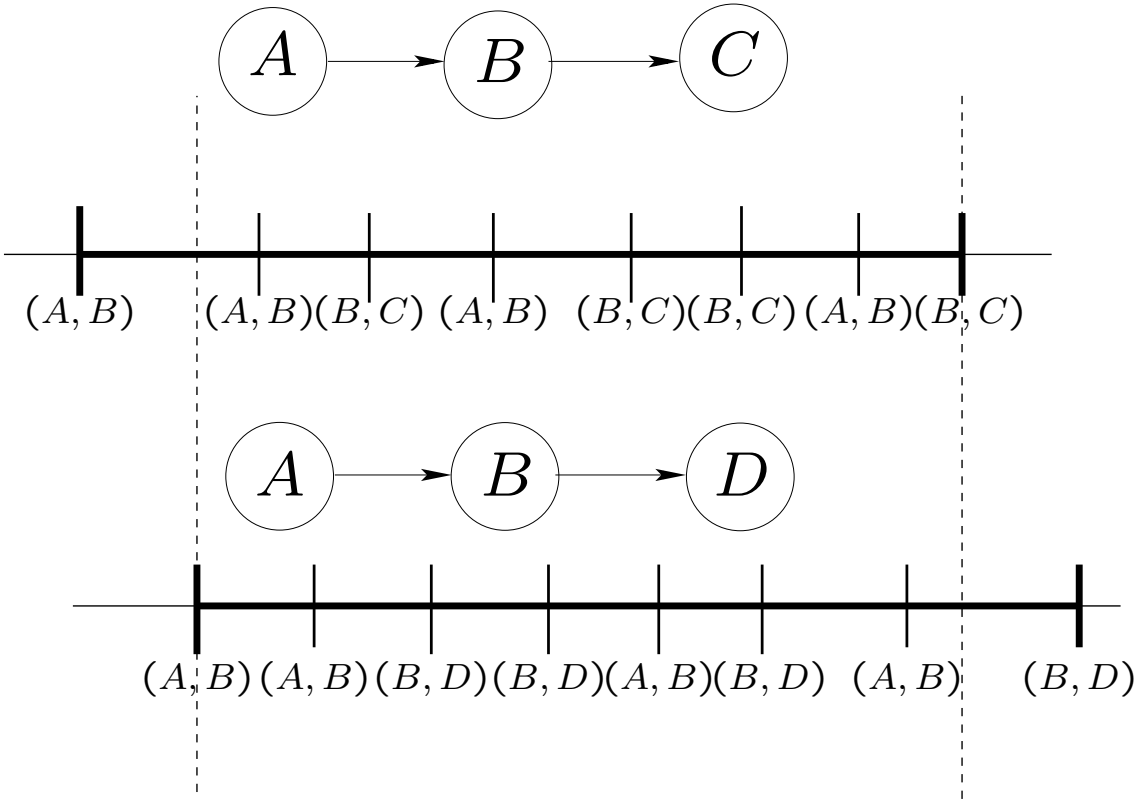
1. Given $A \to B \to C$: finds max matching in linear time $O(n + m)$.

2. All significant triples: $O(deg \cdot \|Data\|)$.

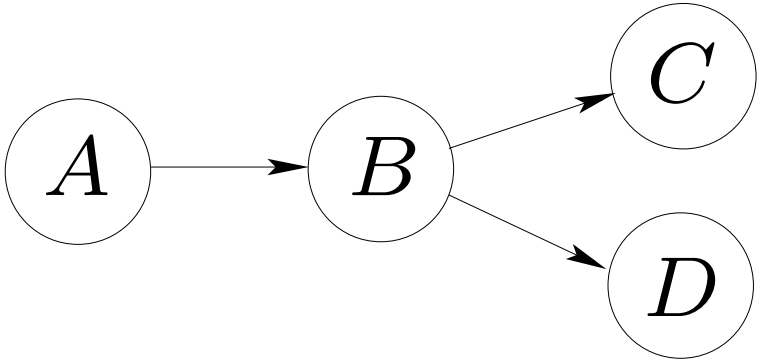


(deg = max numer of receivers per sender.)

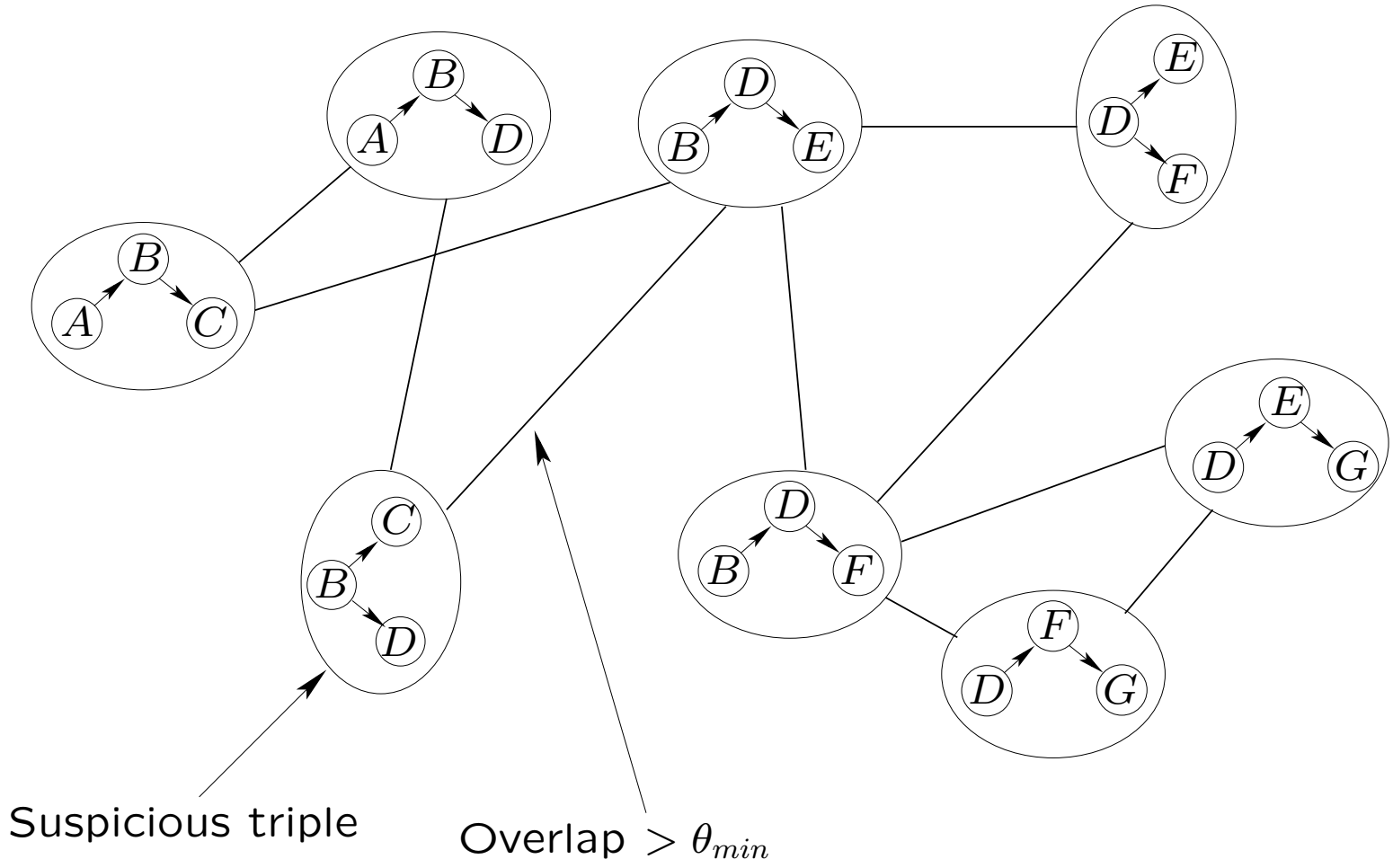
Building Larger Groups Using Overlaps



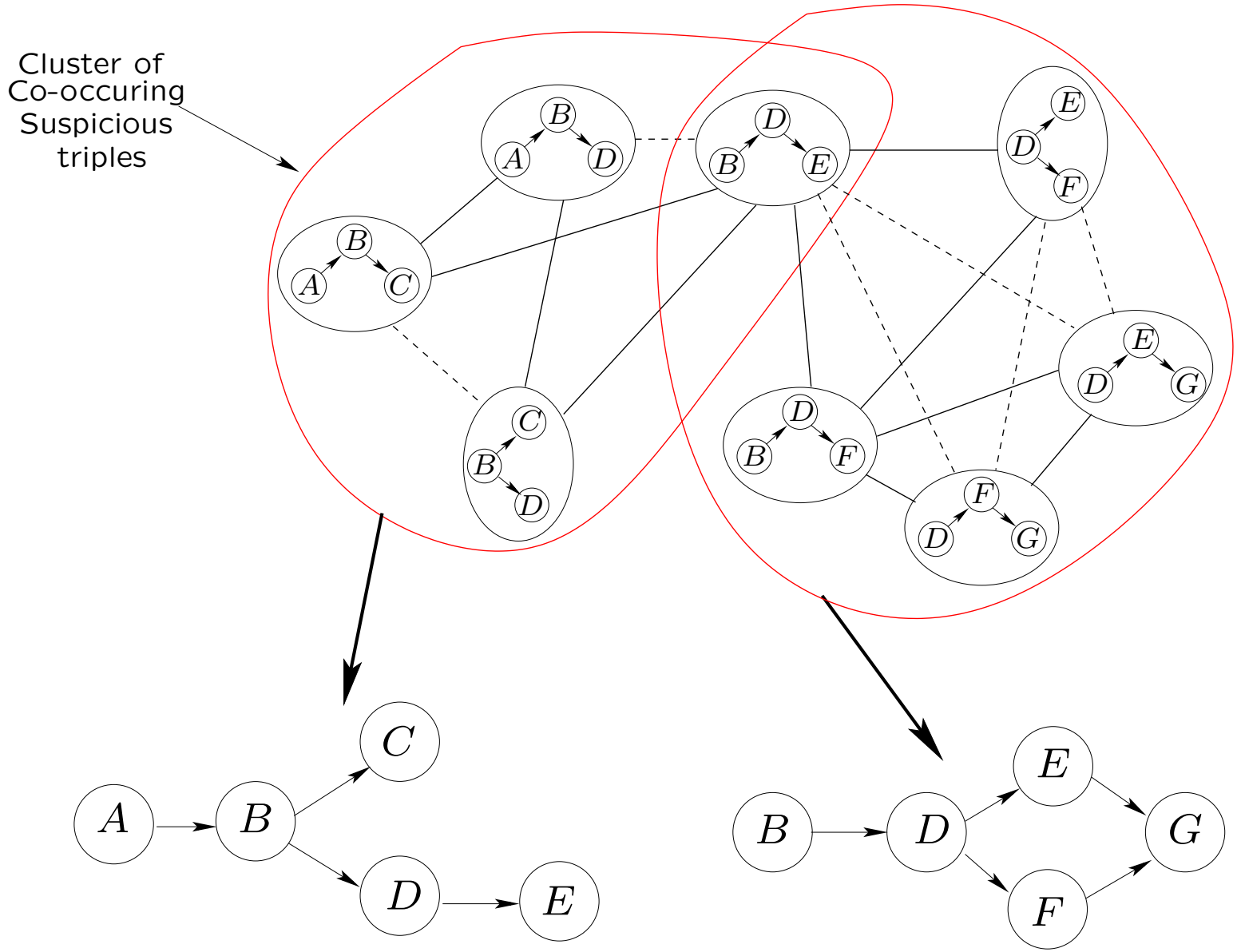
If **overlap** $> \theta_{min}$, then:



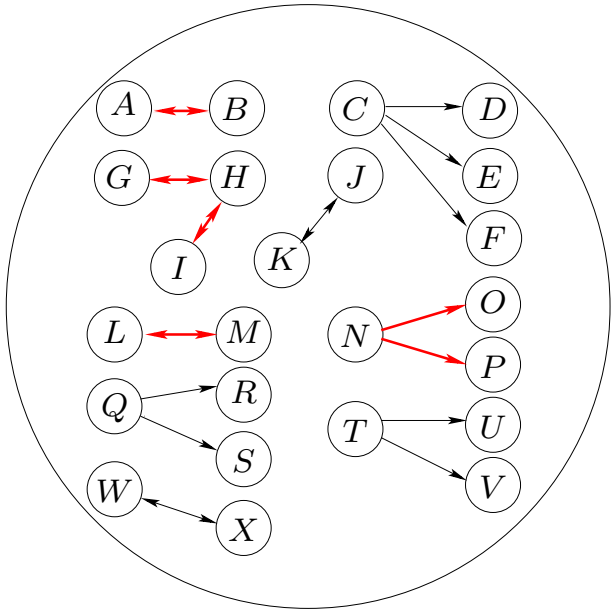
Overlap Graph



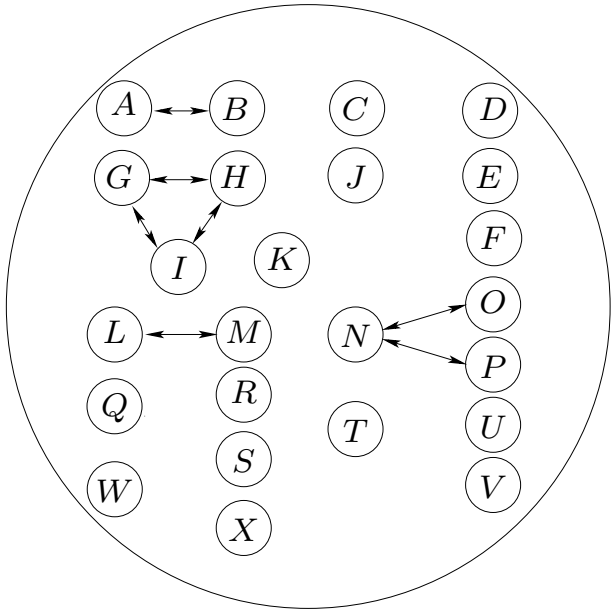
Clustering



Weblog Results (Feb. 2006)



Discovered Weblog Groups



Published Friendship Graph

Probability of a random edge $P_{edge} = 0.0008\%$

Found Group = 2.5%

Recap

1. **Find triples** - use Max Matching (linear time).
   - Can be extended to arbitrary sized chains and siblings.

2. Determine **statistical significance threshold** for triples.

3. Build **overlap graph** of **suspicious** triples.

4. **Cluster overlap graph** to get groups of concurrent triples.
   - group structure must be consistent with triples.

5. **Track evolution:** group and internal communication flow.

This is Ongoing Work ...

1. General tree/DAG querying algorithm.

2. Using bigger Chains and Siblings as base (instead of triples).

3. Exact algorithm for finding general trees/DAGs.

4. Probabilistic propagation delay functions.

Thank You!

http://www.cs.rpi.edu/∼magdon/

References

[BGMW04]. Baumes, J., Goldberg, M., Magdon-Ismail, M., Wallace, W.: Discovering hidden groups in communication networks. ISI 2004.

[CSCC04]. Capocci, A., Servedio, V.D.P., Caldarelli, G., Colaiori, F.: Detecting communities in large networks. WAW 2004.

[N03]. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 2003.

[GN02]. Girvan, M., Newman, M.E.J.: Community structure in social and biological net- works. Proc. Natl. Acad. Sci. 99, 2002.

[MGWS03]. Magdon-Ismail, M., Goldberg, M., Wallace, W., Siebecker, D.: Locating hidden groups in communication networks using Hidden Markov Models. ISI 2003.

[BGKMP05]. Baumes, J., Goldberg, M., Krishnamoorthy, M., Magdon-Ismail, M., Preston, N.: Finding comminities by clustering a graph into overlapping subgraphs. IADIS Applied Computing 2005.

[BGM05]. Baumes, J., Goldberg, M., Magdon-Ismail, M.: Efficient identification of overlapping communities. ISI 2005.

[BGHMWZ06]. Baumes, J., Goldberg, M., Hayvanovych, M., Magdon-Ismail, M., Wallace, W., Zaki, M.: Algorithms For Finding Hidden Group Structure From Communication Data. Submitted to IEEE TKDE 2006.