SCS CMU

Large Graph Mining: Patterns, Tools and Case Studies

Christos Faloutsos
Hanghang Tong
CMU

SCS CMU

Outline

- Part 1: Patterns
- Part 2: Matrix and Tensor Tools
- ➡ Part 3: Proximity
- Part 4: Case Studies

SCS CMU

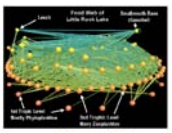
Part 3: Proximity on Graphs

-Definitions, Fast Solutions, and Applications

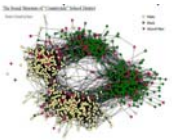
SCS CMU

Recap: Graphs are everywhere!

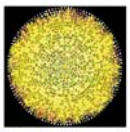
Internet Map
[lumeta.com]



Food Web
[Martinez '91]



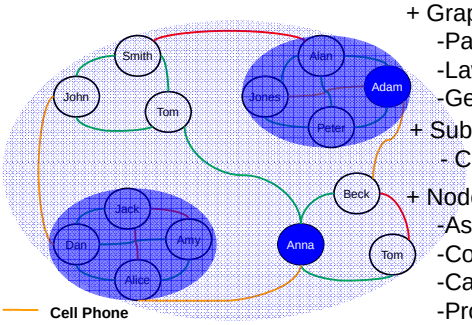
Friendship Network
[Moody '01]



Protein Interactions
[genomebiology.com]

SCS CMU

Graph Mining: Big Picture

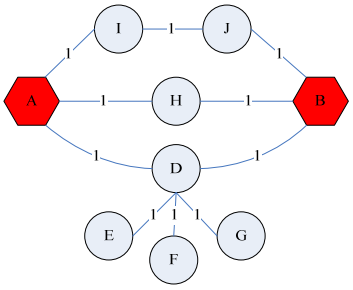


+ Graph Level
  -Patterns
  -Laws
  -Generators
+ Subgraph Level
  - Community
+ Node Level
  -Association
  -Correlation
  -Causality
  -Proximity

We are here!

SCS CMU

Proximity on Graph: What?



a.k.a Relevance, Closeness, 'Similarity'…

SCS CMU

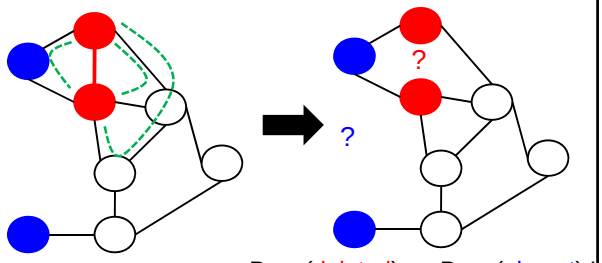
Proximity on Graphs: Why?

- A1: Link Prediction [Liben-Nowell+], [Tong+]
- A2: Neighborhood Search [Sun+]
- A3: Image Captioning [Pan+]
- A4: Conn. Subgraph [Faloutsos+], [Tong+], [Koren+]
- A5: Best Effort Pattern match [Tong+]
- A6: Proximity Tracking [Tong+]
- A7: Mining Time [Chakrabarti+]
- A8: Interactive Querying and Summarization [Tong+]
- Many more…

Will return to this later

SCS CMU

A1: Link Prediction

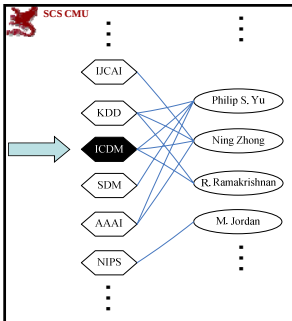


Prox (deleted) >> Prox (absent) !

Footnote:
- Red pair: ``deleted'';
- Blue pair: ``absent''

SCS CMU

A2: Neighborhood Search on graphs



Conference Author

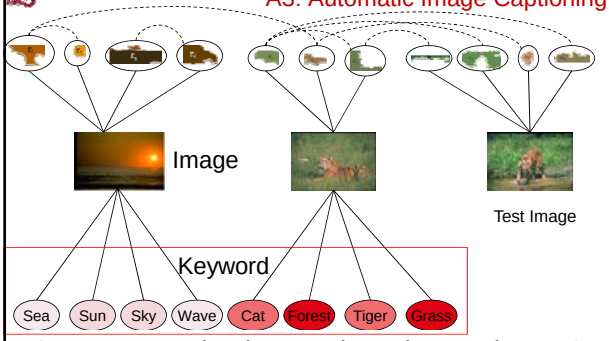
Q: what is most related conference to ICDM?

A: Proximity! [Sun+ ICDM2005]

SCS CMU

Region

A3: Automatic Image Captioning

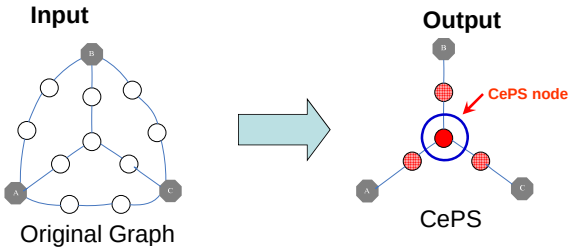


Q: How to assign keywords to the test image?

A: Proximity! [Pan+ 2004]

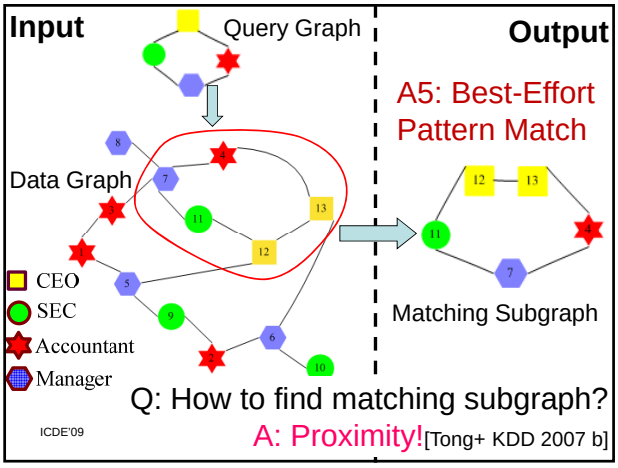
SCS CMU

A4: Center-Piece Subgraph(CePS)



Q: How to find hub for the black nodes?

A: Proximity! [Tong+ KDD 2006]



Q: How to find matching subgraph?

A: Proximity![Tong+ KDD 2007 b]

SCS CMU

Outline: Part 3

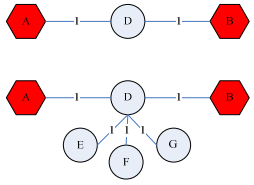
- Motivation
- Definitions
- Fast Solutions
- Applications
- Conclusion

- Basic: RWR
- Variants
- Properties
- Generalizations

SCS CMU

Some ``bad'' proximities

Why not shortest path?

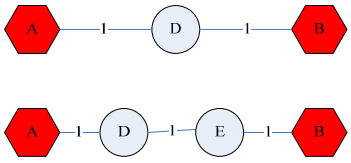


'pizza delivery guy' problem

SCS CMU

Some ``bad'' proximities

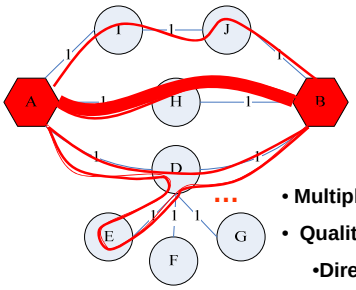
Why not max. netflow?



No punishment on long paths

SCS CMU

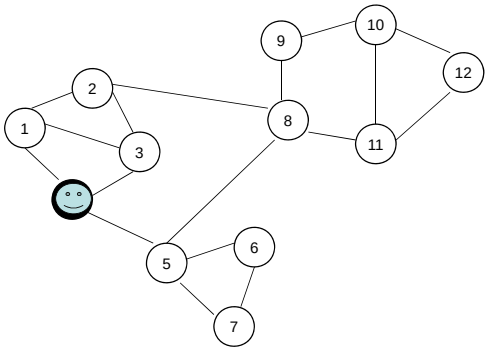
What is a ``good'' Proximity?



- **Multiple Connections**
- **Quality of connection**
  - **•Direct & In-directed Conns**
  - **•Length, Degree, Weight…**

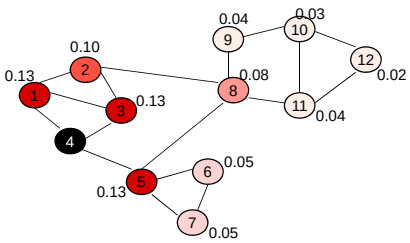
SCS CMU

Random walk with restart



SCS CMU

Random walk with restart



Nearby nodes, higher scores

More red, more relevant

	Node 4
Node 1	0.13
Node 2	0.10
Node 3	0.13
Node 4	0.22
Node 5	0.13
Node 6	0.05
Node 7	0.05
Node 8	0.08
Node 9	0.04
Node 10	0.03
Node 11	0.04
Node 12	0.02

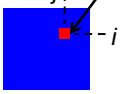
Ranking vector $\vec{r}_4$

SCS CMU

Why RWR is a good score?

$Q(i, j) \propto r_{i,j}$

$Q = (I - c\tilde{W})^{-1} =$



$\tilde{W}$: adjacency matrix.
c: damping factor

$Q = c\,\tilde{W} + c^2\tilde{W}^2 + c^3\tilde{W}^3 + \cdots$

all paths from *i* to *j* with length **1**

all paths from *i* to *j* with length 2

all paths from *i* to *j* with length 3

SCS CMU

Outline: Part 3

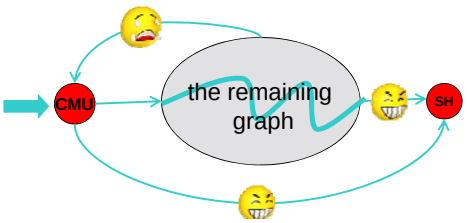
- Motivation
- Definitions
- Fast Solutions
- Applications
- Conclusion

- Basic: RWR
- Variants
- Properties
- Generalizations

SCS CMU

Variant: escape probability

- Define Random Walk (RW) on the graph
- Esc_Prob(CMU→Shanghai)
  - Prob (starting at CMU, reaches Shanghai before returning to CMU)



Esc_Prob = Pr (smile before cry)

SCS CMU

Other Variants

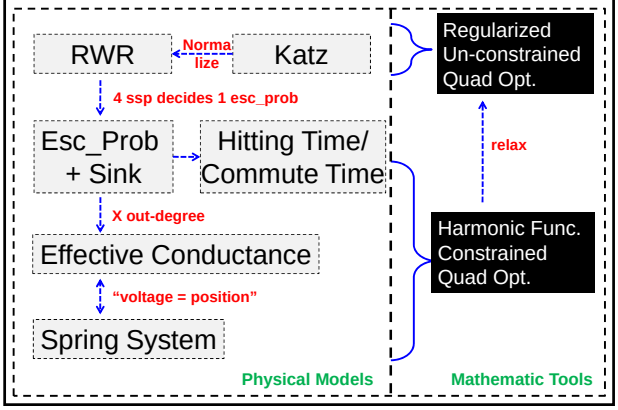
- Other measure by RWs
  - Community Time/Hitting Time [Fouss+]
  - SimRank [Jeh+]

All are "related to" or "similar to" random walk with restart!

  - Spring Systems
- Katz [Katz], [Huang+], [Scholkopf+]
- Matrix-Forest-based Alg [Chobotarev+]

SCS CMU

Overview of different measurements



SCS CMU

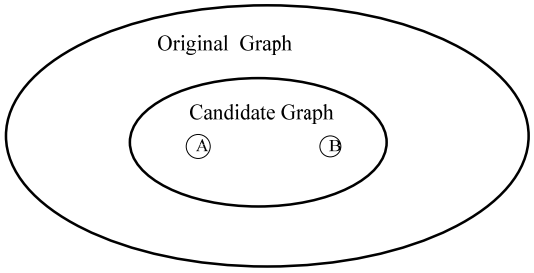
Outline: Part 3

- Motivation
- Definitions
- Fast Solutions
- Applications
- Conclusion

- Basic: RWR
- Variants
- Properties
- Generalizations

SCS CMU

Property #1: Monotonicity

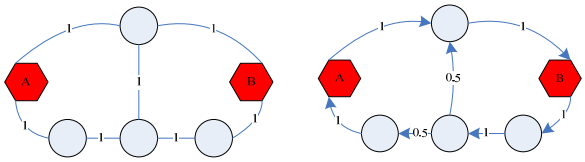


We want: $\text{Prox}^{candi}(a \to b) \leq \text{Prox}^{original}(a \to b)$

A: degree preserving! [Koren+ KDD06][Tong+ KDD07a][Tong+ SDM08]

SCS CMU

Property #2: Asymmetry [Tong+ KDD07 a]



What is Prox between A and B?

What is Prox from A to B?
What is Prox from B to A?

SCS CMU

Asymmetry in un-directed graphs

- Hanghang's # 1 Conf. is ICDE
- The #1 author of ICDE is ...

SCS CMU

Outline: Part 3

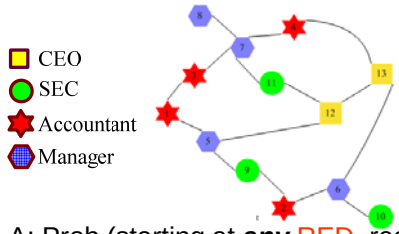
- Motivation
- Definitions
- Fast Solutions
- Applications
- Conclusion

- Basic: RWR
- Variants
- Properties
- Generalizations

SCS CMU

Group Proximity [Tong+ KDD07 a]

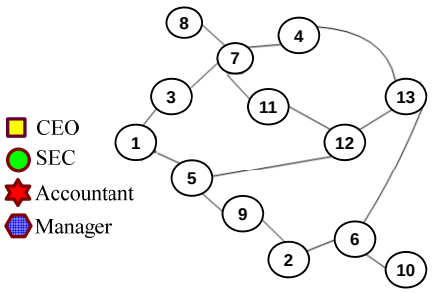
- Q: *How close are Accountants to SECs?*



- A: Prob (starting at ***any*** RED, reaches ***any*** GREEN before touching ***any*** RED again)

SCS CMU

Proximity on Attributed Graphs [Tong+ KDD07 b]

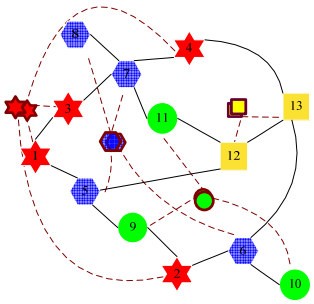


What is the proximity from node 7 to 5?

If we know that…

SCS CMU

A: Augmented graphs [Tong+ KDD07 b]



SCS CMU

More on Generalizations

- Attributes on edges [Chakrabarti+ KDD 06]
- Proximity w/ Time
  - [Minkov+], [Tong+ SDM 2008], [Tong+ CIKM 2008]
- Proximity w/ Side Information [Tong+ 2008]
- …

SCS CMU

Summary of Proximity Definitions

- Goal: Summarize multiple … relationship
- Solutions
  - Basic: Random Walk with Restart
    - [Pan+ 2004][Sun+ 2006][Tong+ 2006]
  - Properties: Asymmetry, monotonicity
    - [Koren+ 2006][Tong+ 2007] [Tong+ 2008]
  - Variants: Esc_Prob and many others.
    - [Faloutsos+ 2004] [Koren+ 2006][Tong+ 2007]
  - Generalizations: Group Prox, w/ Attr., w/ Time, w/ Side Information
    - [Charkrabarti+ 2006][Tong+ 2007] [Tong+ 2008]

SCS CMU

Outline: Part 3

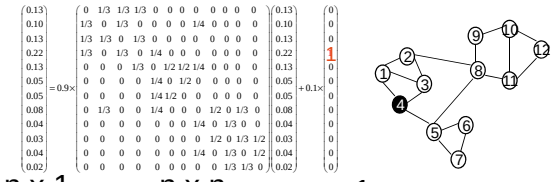
- Motivation
- Definitions
- Fast Solutions
- Applications
- Conclusion

SCS CMU

Computing RWR

$$\vec{r}_i = c\tilde{W}\vec{r}_i + (1-c)\vec{e}_i$$

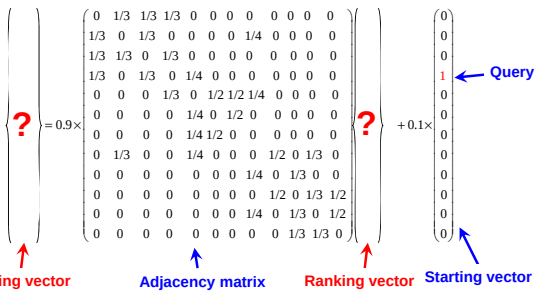
Ranking vector | Adjacency matrix | Restart p | Starting vector



n x 1 n x n n x 1

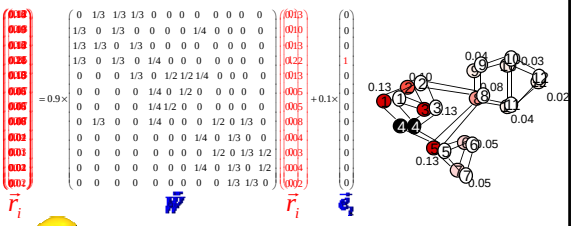
SCS CMU

Q: Given query *i*, how to solve it?



SCS CMU

OntheFly: $\vec{r}_i[t+1] = c\tilde{W}\vec{r}_i[t] + (1-c)\vec{e}_i$



No pre-computation/ light storage

Slow on-line response O(mE)

SCS CMU

B_Lin: Basic Idea

[Tong+ ICDM 2006]

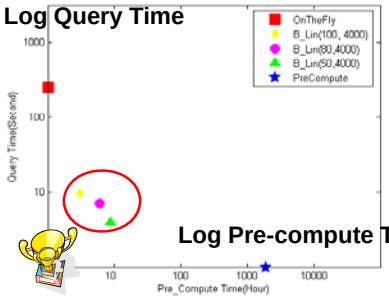
Find Community

Combine

Fix the remaining

SCS CMU

Query Time vs. Pre-Compute Time



- **Quality: 90%+**
- **On-line:**
  - **Up to 150x speedup**
- **Pre-computation:**
  - **Two orders saving**

SCS CMU

Summary of Fast Solutions

- B_Lin: one large linear system [Tong+ ICDM06]
- BB_Lin: the intrinsic complexity is small
  - [Tong+ ICDM06]
- FastUpdate: dynamic linear system [Tong+ SDM08]
- FastAllDAP: multiple linear systems [Tong+ KDD07 a]
- MT3: multiple-resolution analysis [Tong+ CIKM 2008]
- Fast-ProSIN: incorporate on-line user's feedback
  - [Tong+ ICDM 2008]

More Details: Off-line!

SCS CMU

Outline: Part 3

- Motivation
- Definitions
- Fast Solutions
- Applications
- Conclusion

SCS CMU

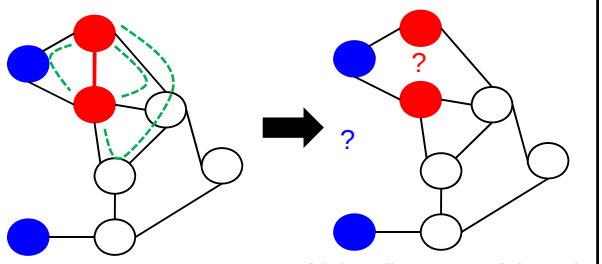
Outline: Part 3

- Motivation
- Definitions
- Fast Solutions
- Applications
- Conclusion

→ Link Prediction & +
- Ranking Related Tasks
- User Specific Patterns
- Time Related Tasks
- Interaction w/ Users

SCS CMU

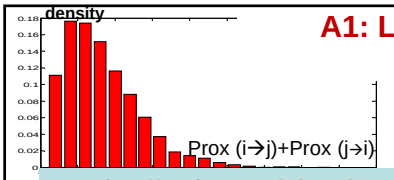
A1: Link Prediction - existence



Prox (deleted) >> Prox (absent) !

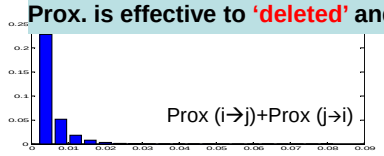
Footnote:
- Red pair: ``deleted'';
- Blue pair: ``absent''

A1: Link Prediction-existence



Prox. Hist. for a set of deleted links

Prox. is effective to 'deleted' and absent edges!



Prox. Hist. for a set of absent links

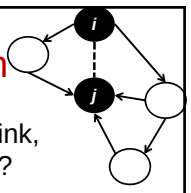
Q: How to predict the existence of the link?

A: Proximity! [Liben-Nowell + 2003]

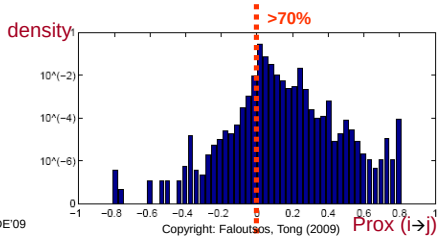
SCS CMU

Link Prediction: direction

[Tong+ KDD 07 a]



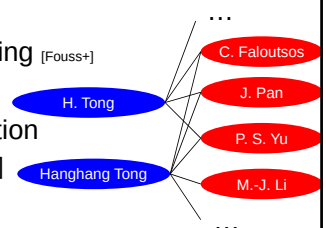
- Q: Given the existence of the link, what is the direction of the link?
- A: Compare prox(i→j) and prox(j→i)



SCS CMU

Beyond Link Prediction

- Collaborative Filtering [Fouss+]
- Name Disambiguation
  - [Minkov+ SIGIR 06]
- Anomaly Nodes/Edges
  - 'a' is abnormal if the neighborhood of 'a' is so different
  - [Sun+ ICDM 2005]



SCS CMU

Outline: Part 3

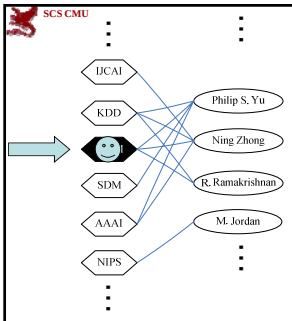
- Motivation
- Definitions
- Fast Solutions
- Applications
- Conclusion

→ Ranking Related Tasks

- Link Prediction & +
- User Specific Patterns
- Time Related Tasks
- Interaction w/ Users

SCS CMU

A2: Neighborhood Search on graphs



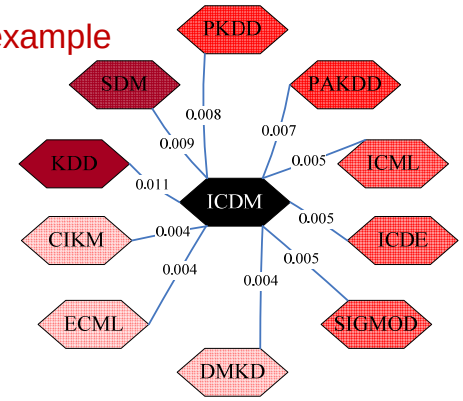
Conference Author

Q: what is most related conference to ICDM?

A: Proximity! [Sun+ ICDM2005]

SCS CMU

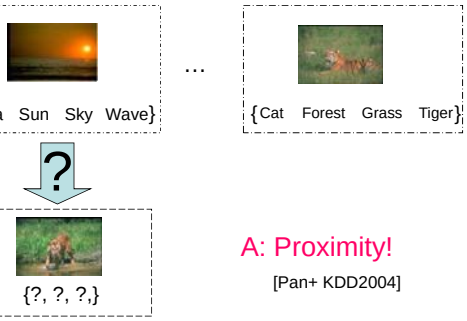
NF: example



SCS CMU

A3: (gCaP) Automatic Image Captioning

- Q

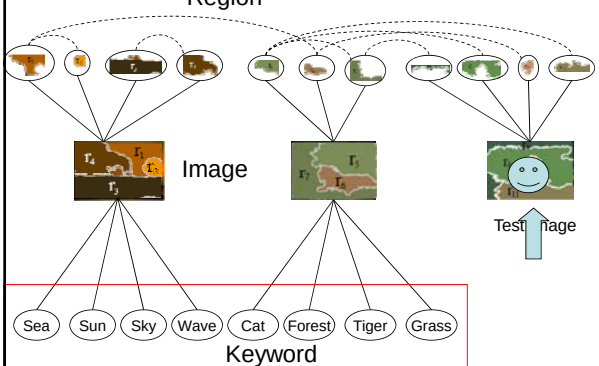


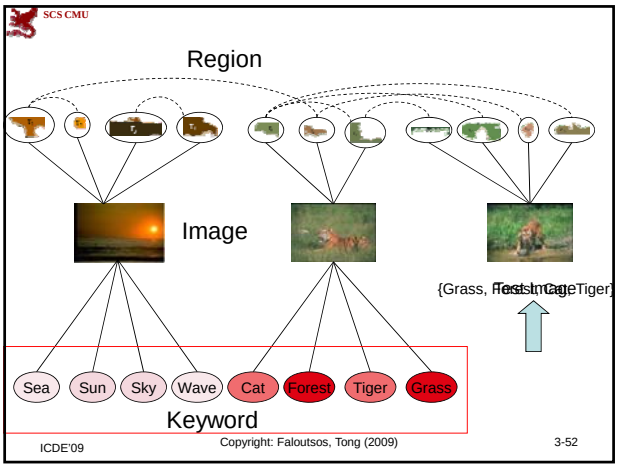
A: Proximity!

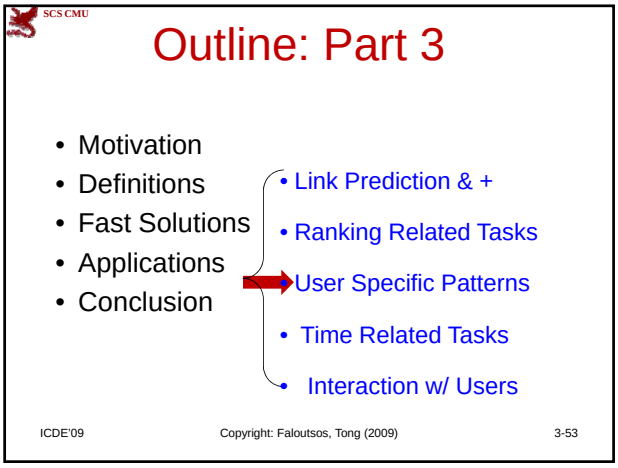
[Pan+ KDD2004]

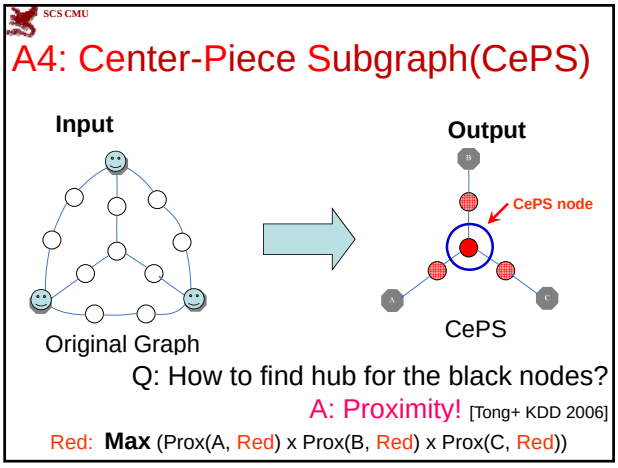
SCS CMU

Region




SCS CMU
Region
Image
{Grass, Forest, Cat, Tiger}
Sea
Sun
Sky
Wave
Cat
Forest
Tiger
Grass
Keyword
ICDE'09
Copyright: Faloutsos, Tong (2009)
3-52


SCS CMU
Outline: Part 3
• Motivation
• Definitions
• Fast Solutions
• Applications
• Conclusion
• Link Prediction & +
• Ranking Related Tasks
• User Specific Patterns
• Time Related Tasks
• Interaction w/ Users
ICDE'09
Copyright: Faloutsos, Tong (2009)
3-53

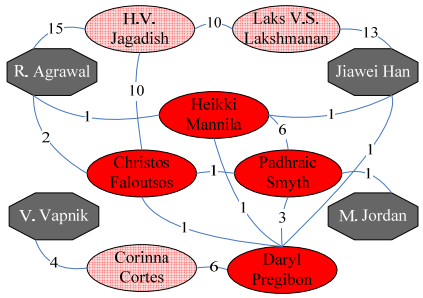

SCS CMU
A4: Center-Piece Subgraph(CePS)
Input
Output
CePS node
Original Graph
CePS
Q: How to find hub for the black nodes?
A: Proximity! [Tong+ KDD 2006]
Red: Max (Prox(A, Red) x Prox(B, Red) x Prox(C, Red))

SCS CMU

CePS: Example

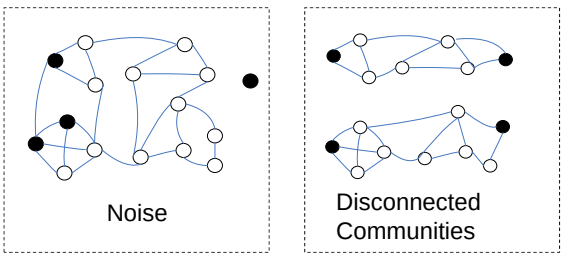
SCS CMU

CePS: Example

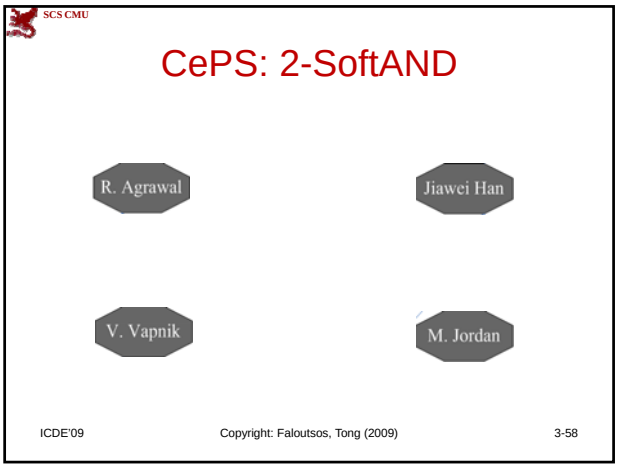


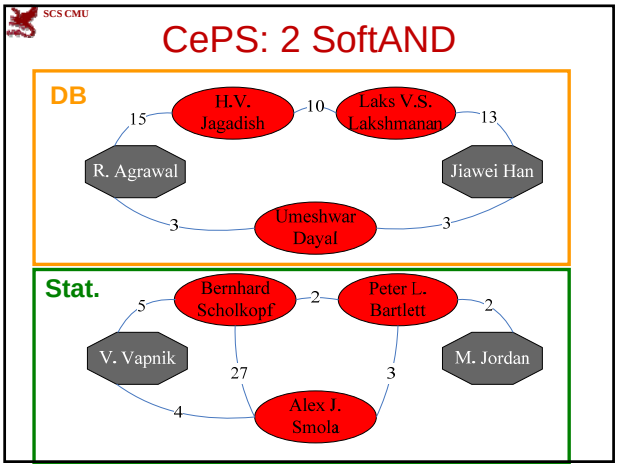
SCS CMU

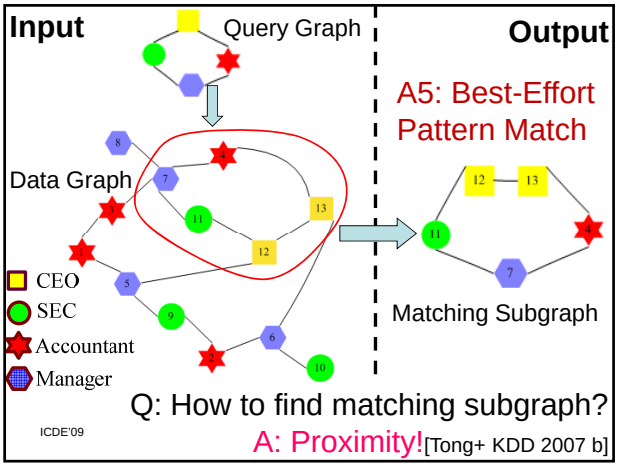
K_SoftAnd: Relaxation of AND



Asking AND query? ➔ No Answer!


SCS CMU
CePS: 2-SoftAND
R. Agrawal
Jiawei Han
V. Vapnik
M. Jordan
ICDE'09
Copyright: Faloutsos, Tong (2009)
3-58

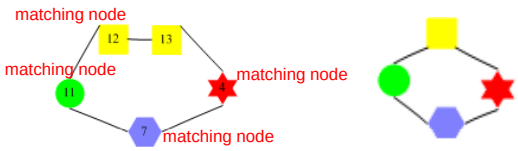

SCS CMU
CePS: 2 SoftAND
DB
H.V. Jagadish
Laks V.S. Lakshmanan
R. Agrawal
Jiawei Han
Umeshwar Dayal
15
10
13
3
3
Stat.
Bernhard Scholkopf
Peter L. Bartlett
V. Vapnik
M. Jordan
Alex J. Smola
5
2
2
27
3
4


Input
Query Graph
Output
A5: Best-Effort Pattern Match
Data Graph
Matching Subgraph
CEO
SEC
Accountant
Manager
Q: How to find matching subgraph?
A: Proximity! [Tong+ KDD 2007 b]
ICDE'09

SCS CMU

G-Ray: How to?

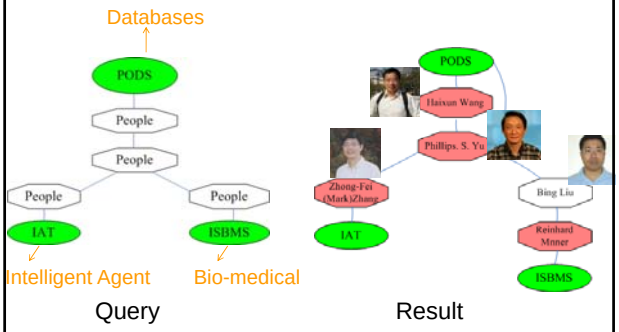
details



Goodness = Prox (12, 4) x Prox (4, 12) x
Prox (7, 4) x Prox (4, 7) x
Prox (11, 7) x Prox (7, 11) x
Prox (12, 11) x Prox (11, 12)

SCS CMU

Effectiveness: star-query



Query Result

SCS CMU

Outline: Part 3

- Motivation
- Definitions
- Fast Solutions
- Applications
- Conclusion

- Link Prediction & +
- Ranking Related Tasks
- User Specific Patterns
- Time Related Tasks
- Interaction w/ Users

SCS CMU

Challenge

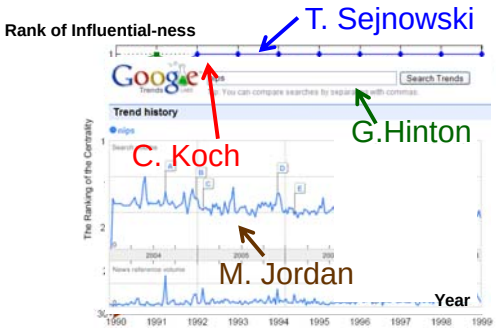
- Graphs are evolving over time!
  - –New nodes/edges show up;
  - –Existing nodes/edges die out;
  - –Edge weights change…

Q: How to generalize everything?

A: Track Proximity! [Tong+ SDM 2008]

SCS CMU

pTrack/cTrack: Trend analysis on graph level



SCS CMU

A6: (pTrack) Problem Definition

- **[Given]**
  - – (1) a large, skewed time-evolving bipartite graphs,
  - – (2) the query nodes of interest
- **[Track]**
  - – (1) top-k most related nodes for each query node at each time step t;
  - – (2) the proximity score (or rank of proximity) between any two query nodes at each time step t

SCS CMU

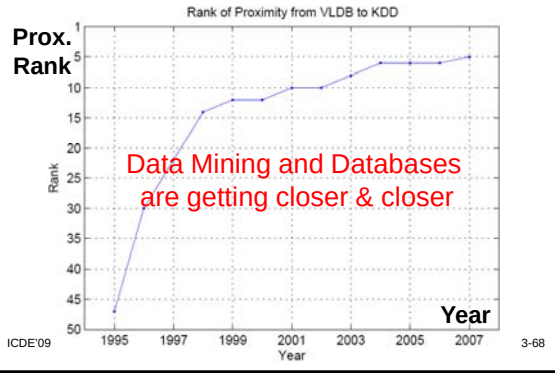
pTrack: Philip S. Yu's Top-5 conferences up to each year

1992	1997	2002	2007
ICDE	CIKM	KDD	ICDM
ICDCS	ICDCS	SIGMOD	KDD
SIGMETRICS	ICDE	ICDM	ICDE
PDIS	SIGMETRICS	CIKM	SDM
VLDB	ICMCS	ICDCS	VLDB

DBLP: (Au. x Conf.)
- 400k aus,
- 3.5k confs
- 20 yrs

SCS CMU

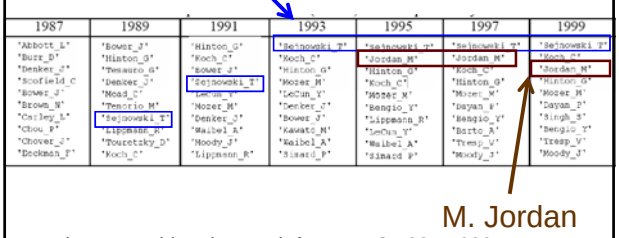
KDD's Rank wrt. VLDB over years



SCS CMU

cTrack:10 most influential authors in NIPS community up to each year

T. Sejnowski



M. Jordan

Author-paper bipartite graph from NIPS 1987-1999.
3k. 1740 papers, 2037 authors,
spreading over 13 years

SCS CMU

A7: (T3/MT3) How to mine time in some complex settings?

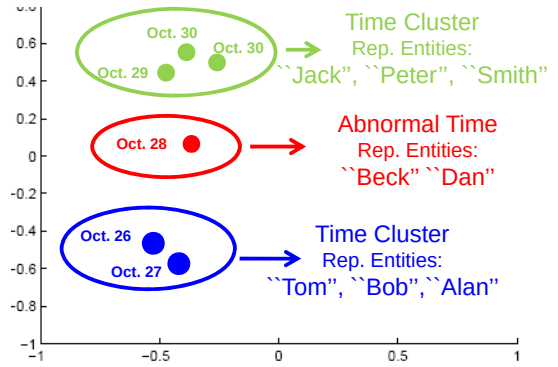
SCS CMU

A Motivating Example: Inputs

Time	Event (e.g., Session)	Entity

SCS CMU

A Motivating Example: Outputs



SCS CMU

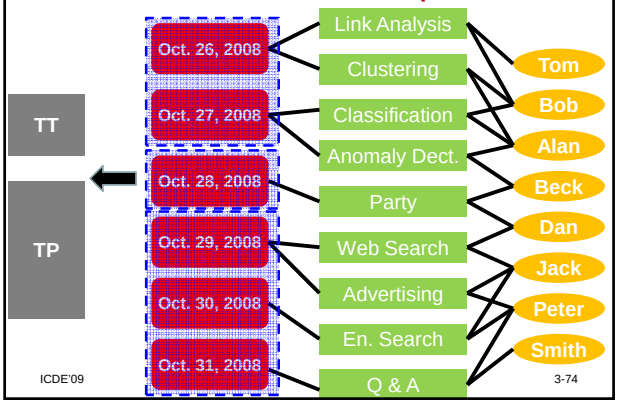
Problem Definitions:

(How to Mine Time in such complex context)

- Given datasets collected at different time stamps;
- Find
  - Q1: Time Cluster
  - Q2: Abnormal Time stamp
  - Q3: Interpretations
  - Q4: Right time granularity

SCS CMU

T3: Basic Idea – Graph + RWR



SCS CMU

Data Sets

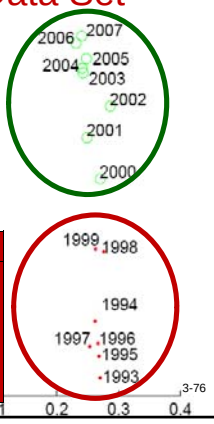
- CIKM: from CIKM proceedings
  - Time: Publication years (1993-2007, 15)
  - Event: Paper (952)
  - Entities: Authors (1895) & Sessions (279)
  - Attribute: Keyword (158)

SCS CMU

T3 on `CIKM' Data Set

Rep. Authors	Rep. Keywords
James. P. Callan	Web
W. Bruce Croft	Cluster
James Allan	Classification
Philip S. Yu	XML
George Karypis	Language
Charles Clarke	Stream

Rep. Authors	Rep. Keywords
Elke Rundensteiner	Knowledge
Daniel Miranker	System
Andreas Henrich	Unstructured
Il-Yeol Song	Rule
Scott B Huffman	Object-oriented
Robert J. Hall	Deductive



3-76

SCS CMU

Outline: Part 3

- Motivation
- Definitions
- Fast Solutions
- Applications
- Conclusion

- Link Prediction & +
- Ranking Related Tasks
- User Specific Patterns
- Time Related Tasks

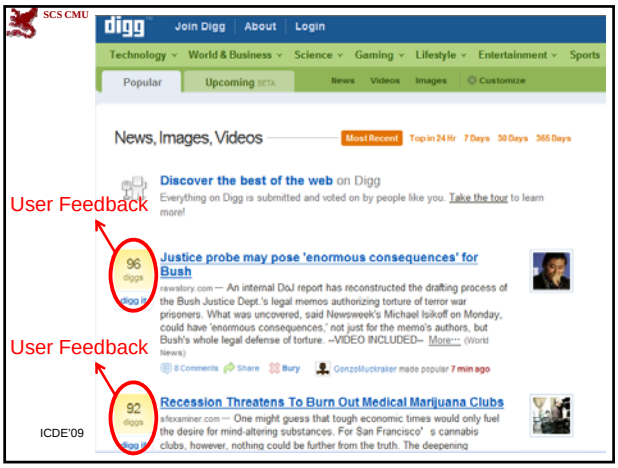
→ Interaction w/ Users

ICDE'09 Copyright: Faloutsos, Tong (2009) 3-77

Interaction with Users

78

SCS CMU



ICDE'09

SCS CMU

Interaction with Users (Querying)

[Tong+ ICDM08]

Initial Results
'ICDM' 'ICML' 'SDM' 'VLDB' 'ICDE' 'SIGMOD' 'NIPS' 'PKDD' 'IJCAI' 'PAKDD'

what are most related conferences wrt KDD?

(DBLP author-conference bipartite graph)

80

SCS CMU

Interaction with Users (Querying)

[Tong+ ICDM08]

Initial Results
'ICDM' 'ICML' X 'SDM' 'VLDB' 'ICDE' 'SIGMOD' 'NIPS' X 'PKDD' 'IJCAI' X 'PAKDD'
two main sub-communities in KDD: DBs (green) vs. Stat (Red)

what are most related conferences wrt KDD?

(DBLP author-conference bipartite graph)

81

SCS CMU

Interaction with Users (Querying)

[Tong+ ICDM08]

Initial Results	No to `ICML'
'ICDM' 'ICML' X 'SDM' 'VLDB' 'ICDE' 'SIGMOD' 'NIPS' X 'PKDD' 'IJCAI' X 'PAKDD'	'ICDM' 'SDM' 'PKDD' 'ICDE' 'VLDB' 'SIGMOD' 'PAKDD' 'CIKM' 'SIGIR' ✓ 'WWW'
two main sub-communities in KDD: DBs (green) vs. Stat (Red)	Negative feedback on ICML will exclude other stats confs (NIPS, IJCAI)

what are most related conferences wrt KDD?

(DBLP author-conference bipartite graph)

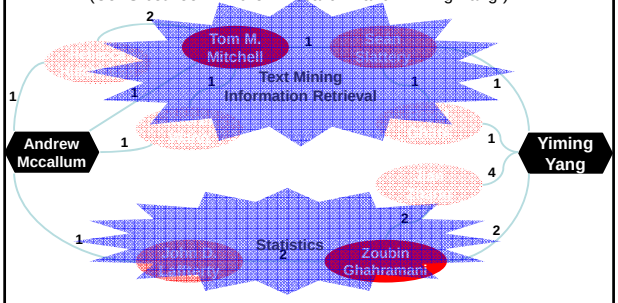
82

SCS CMU

Interaction with Users (Summarization)

[Tong+ ICDM08]

(CePS between "Andrew Mccallum" and "Yiming Yang")



There are two main connections between "Mccallum" and "Yang"

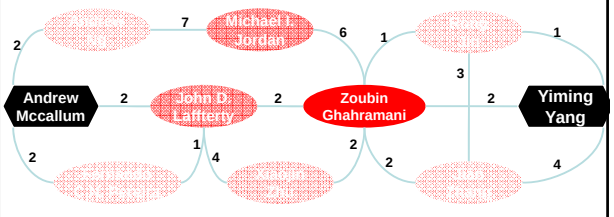
3-83

SCS CMU

Interaction with Users (Summarization)

[Tong+ ICDM08]

(CePS between "Andrew Mccallum" and "Yiming Yang", but avoiding ''Tom M. Mitchell)



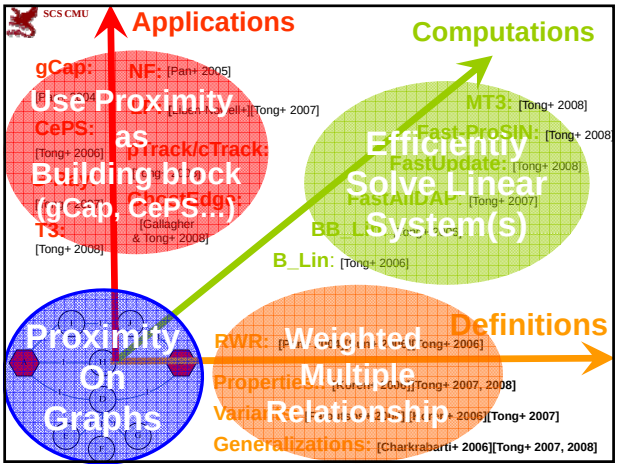
The feedback guides to avoid the entire 'Text' connection, and brings more connections on 'Statistics'

3-84

SCS CMU

More Applications

- Clustering
  - Proximity as input [Ding+ KDD 2007]
- Email management [Minkov+ CEAS 06].
- Business Process Management [Qu+ 2008]
- Ghost Edge
  - Within Network Classification [Gallagher & Tong+ KDD08 b]
- …



SCS CMU

Take-home messages

- Proximity Definitions
  - RWR $\vec{r} = cW\vec{r} + (1-c)\vec{e}$
  - and a lot of variants
- Computations
  - Sherman–Morrison Lemma
  - Fast Incremental Computation
- Applications
  - Proximity as a building block

SCS CMU

References

- L. Page, S. Brin, R. Motwani, & T. Winograd. (1998), The PageRank Citation Ranking: Bringing Order to the Web, Technical report, Stanford Library.
- T.H. Haveliwala. (2002) Topic-Sensitive PageRank. In WWW, 517-526, 2002
- J.Y. Pan, H.J. Yang, C. Faloutsos & P. Duygulu. (2004) Automatic multimedia cross-modal correlation discovery. In KDD, 653-658, 2004.
- C. Faloutsos, K. S. McCurley & A. Tomkins. (2002) Fast discovery of connection subgraphs. In KDD, 118-127, 2004.
- J. Sun, H. Qu, D. Chakrabarti & C. Faloutsos. (2005) Neighborhood Formation and Anomaly Detection in Bipartite Graphs. In ICDM, 418-425, 2005.
- W. Cohen. (2007) Graph Walks and Graphical Models. Draft.

SCS CMU

References

- P. Doyle & J. Snell. (1984) Random walks and electric networks, volume 22. Mathematical Association America, New York.
- Y. Koren, S. C. North, and C. Volinsky. (2006) Measuring and extracting proximity in networks. In KDD, 245–255, 2006.
- A. Agarwal, S. Chakrabarti & S. Aggarwal. (2006) Learning to rank networked entities. In KDD, 14-23, 2006.
- S. Chakrabarti. (2007) Dynamic personalized pagerank in entity-relation graphs. In WWW, 571-580, 2007.
- F. Fouss, A. Pirotte, J.-M. Renders, & M. Saerens. (2007) Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation. IEEE Trans. Knowl. Data Eng. 19(3), 355-369 2007.

SCS CMU

References

- H. Tong & C. Faloutsos. (2006) Center-piece subgraphs: problem definition and fast solutions. In KDD, 404-413, 2006.
- H. Tong, C. Faloutsos, & J.Y. Pan. (2006) Fast Random Walk with Restart and Its Applications. In ICDM, 613-622, 2006.
- H. Tong, Y. Koren, & C. Faloutsos. (2007) Fast direction-aware proximity for graph mining. In KDD, 747-756, 2007.
- H. Tong, B. Gallagher, C. Faloutsos, & T. Eliassi-Rad. (2007) Fast best-effort pattern matching in large attributed graphs. In KDD, 737-746, 2007.
- H. Tong, S. Papadimitriou, P.S. Yu & C. Faloutsos. (2008) Proximity Tracking on Time-Evolving Bipartite Graphs. to appear in SDM 2008.

SCS CMU

References

- B. Gallagher, H. Tong, T. Eliassi-Rad, C. Faloutsos. Using Ghost Edges for Classification in Sparsely Labeled Networks. KDD 2008
- H. Tong, Y. Sakurai, T. Eliassi-Rad, and C. Faloutsos. Fast Mining of Complex Time-Stamped Events CIKM 08
- H. Tong, H. Qu, and H. Jamjoom. Measuring Proximity on Graphs with Side Information. ICDM 2008

