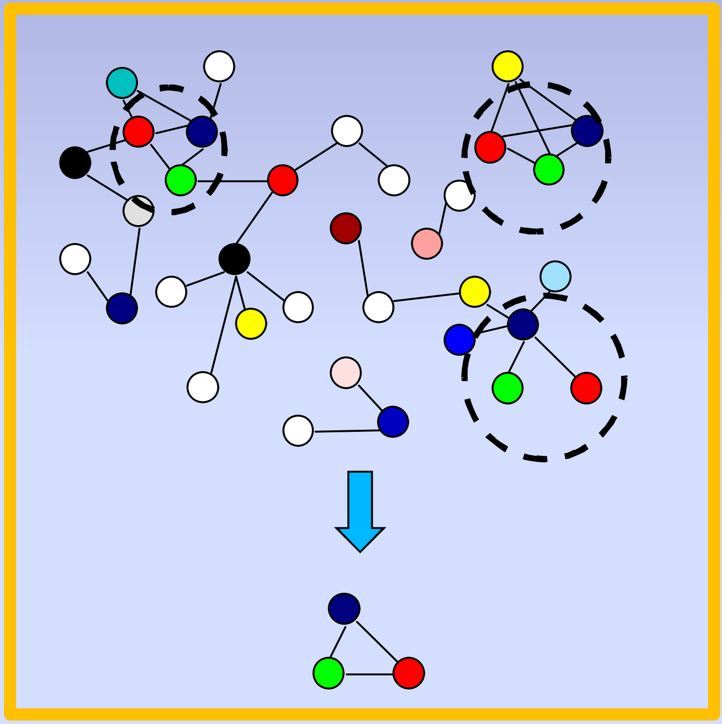
Towards Proximity Pattern Mining in Large Graphs

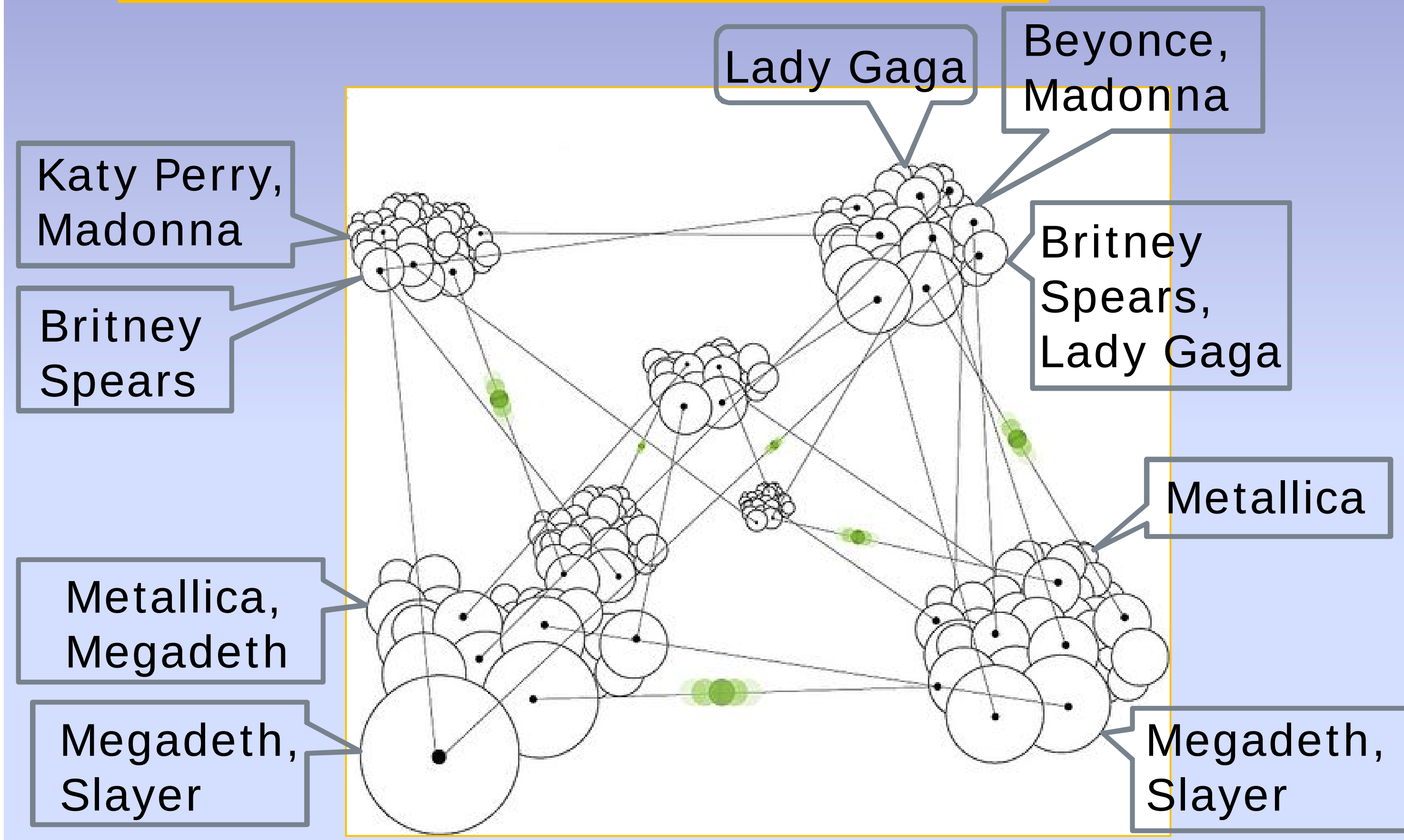


Arijit Khan
Computer Science Department
University of California, Santa Barbara
arijitkhan@cs.ucsb.edu

Xifeng Yan
Computer Science Department
University of California, Santa Barbara
xyan@cs.ucsb.edu

Kun-Lung Wu
IBM T. J. Watson, Hawthorne, NY
klwu@us.ibm.com

Motivation



Homophily in Social Network

Last.FM

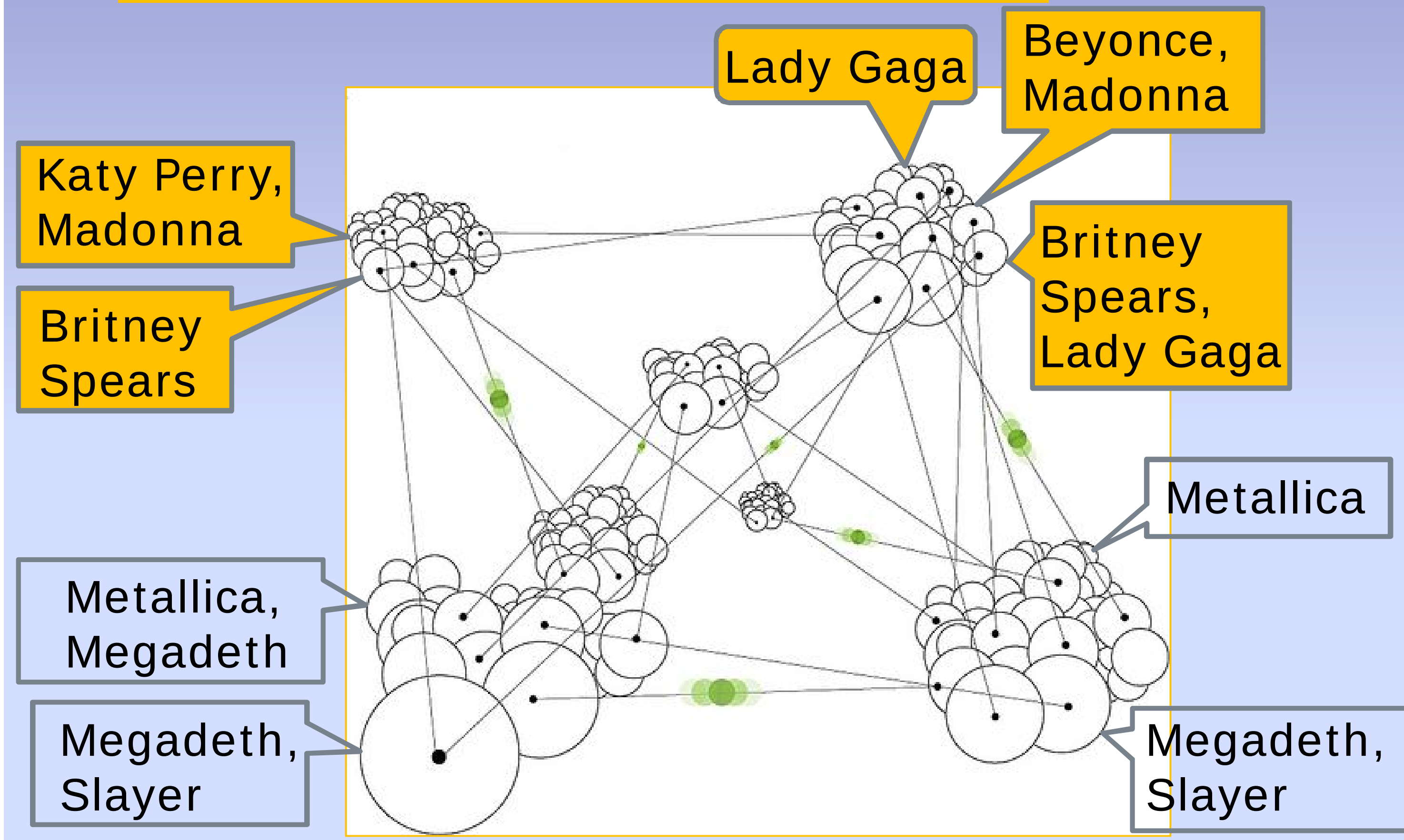
Nodes -> Users

Edges -> Links

List of Musical Bands/ Singers

What are the **related Musical Bands/ Singers** that **co-occur frequently in neighborhood?**

Motivation



Homophily in Social Network

Last.FM

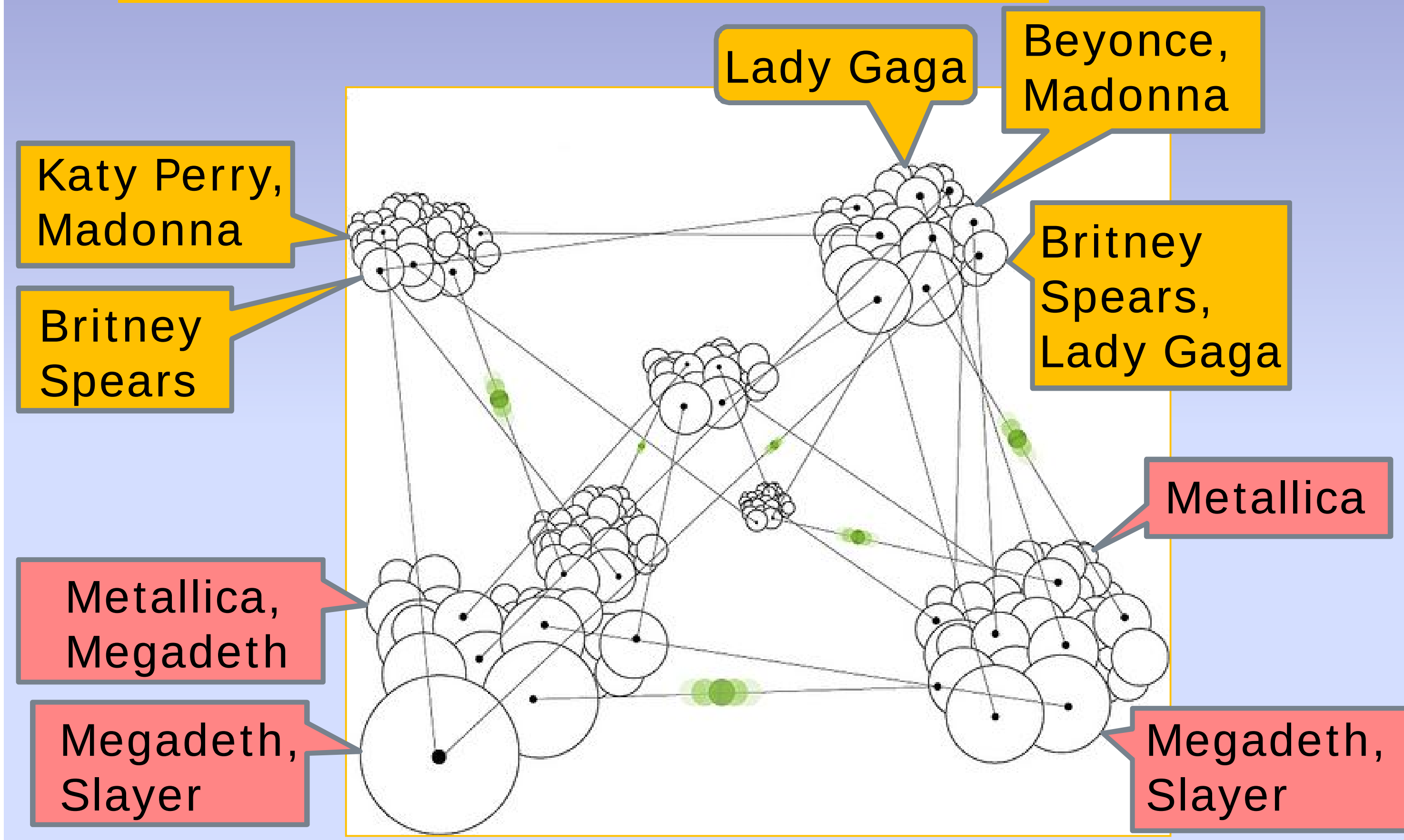
Nodes -> Users

Edges -> Links

List of Musical Bands/ Singers

What are the **related Musical Bands/ Singers** that **co-occur frequently in neighborhood**?

Motivation



Homophily in Social Network

Last.FM

Nodes -> Users

Edges -> Links

List of Musical Bands/ Singers

What are the **related Musical Bands/ Singers** that **co-occur frequently in neighborhood?**

Motivation

Computers in LAN

Computers in Same LAN Attacked by Similar Intrusions

What are **Related Computer Attacks** that **Co-occur Frequently in Neighborhood?**

TFTP_Put, Ping_Flood

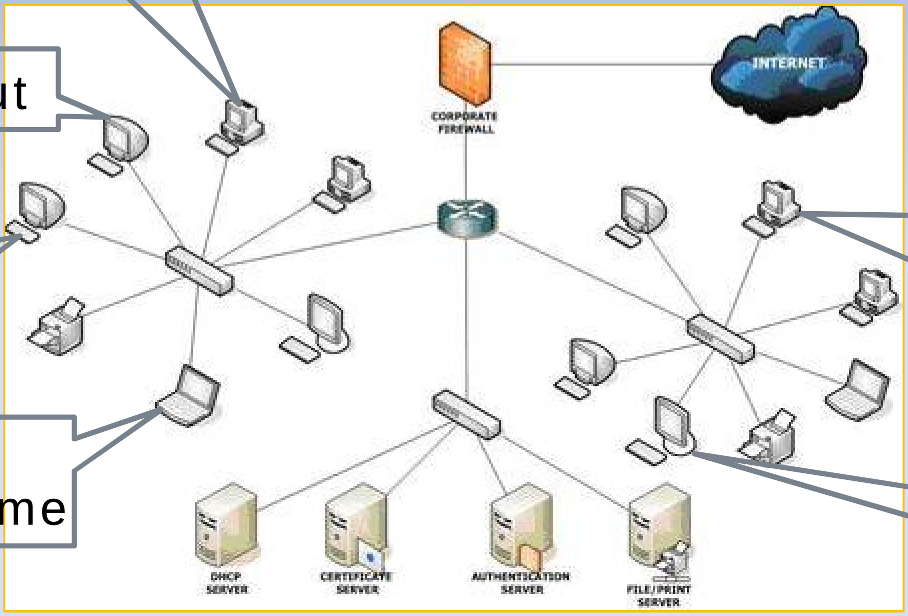
TFTP_Put

TFTP_Put, ICMP_Flood

Audit_TFTP _Get_Filename

SQL _SSRP _Slammer _Worm

SQL _SSRP _ StackBo



Intrusion Network

Motivation

Computers in LAN

Computers in Same LAN Attacked by Similar Intrusions

What are **Related Computer Attacks** that **Co-occur Frequently in Neighborhood?**

TFTP_Put, Ping_Flood

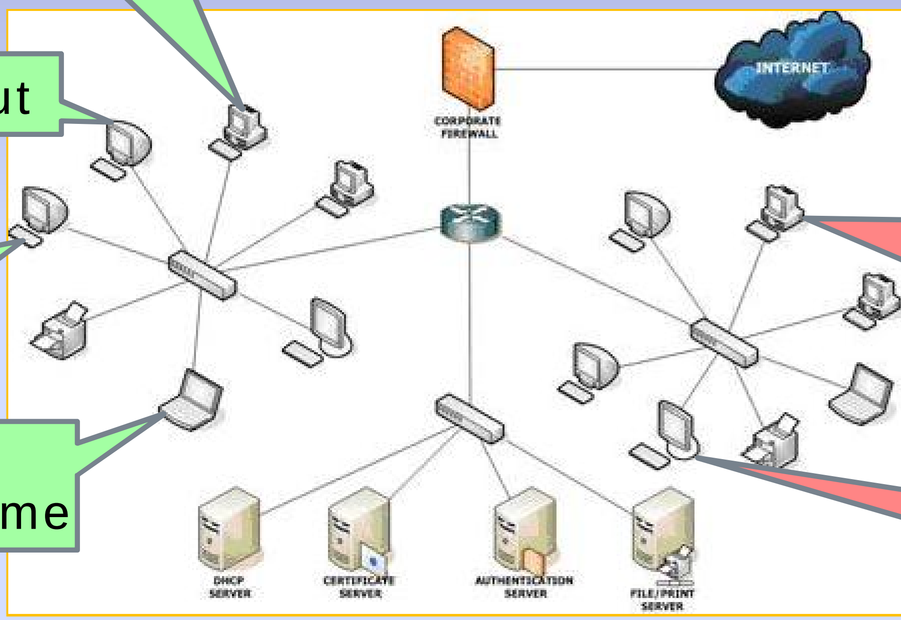
TFTP_Put

TFTP_Put, ICMP_Flood

Audit_TFTP _Get_Filename

SQL _SSRP _Slammer _Worm

SQL _SSRP _ StackBo



Intrusion Network

Roadmap

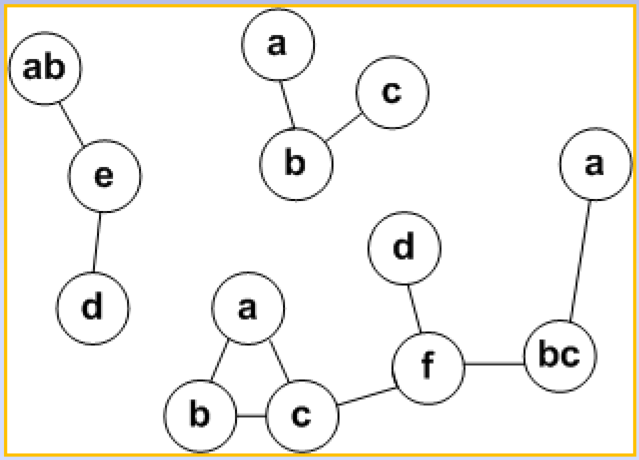
- **Problem Formulation**
  - Problem Definition
  - Preliminaries
- Framework
  - Neighborhood Association Model
  - Information Propagation Model
- Probabilistic Itemset Mining
- Experimental Results
- Conclusion

Problem Definition

Mining Proximity Patterns in Large Graphs.

CHARACTERISTICS

- Proximity
- Frequency



a, b – YES
a, b, c – YES
d, e, f - NO

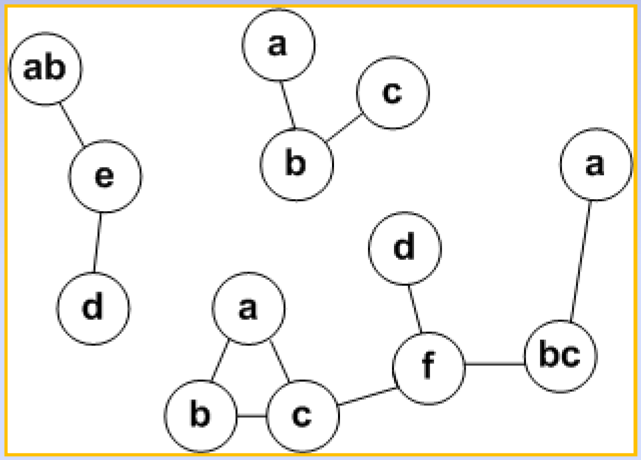
Problem Definition

- Will Frequent Subgraph Mining Work? - NO !!!

- **Flexibility**

- Will Frequent Itemset Mining Work? - NO !!!

- No Notion of Edge in Frequent Itemset Mining



{a, b, c}

Frequent Subgraph – No
Frequent Itemset - No
Proximity Pattern - Yes

Preliminaries

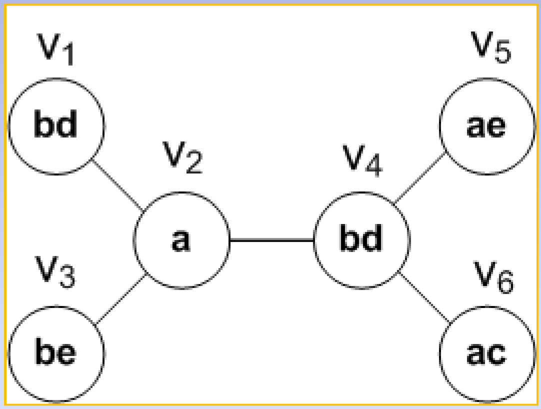
- Labeled Graph $G = (V, E, L)$
- Item Set $I \subseteq L$ is a subset of Labels.
- **SUPPORT**: The support *sup(I)* of an itemset $I \subseteq L$ is the number of transactions in the data set that contain I.
- **DOWNWARD CLOSURE**: For a frequent itemset, all of its subsets are frequent; and thus for an infrequent itemset, all of its superset must be infrequent.

Roadmap

- Problem Formulation
  - Problem Definition
  - Preliminaries
- **Framework**
  - Neighborhood Association Model
  - Information Propagation Model
- Probabilistic Itemset Mining
- Experimental Results
- Conclusion

Neighborhood Association Model

- **EMBEDDING:**
  - $\{v_1, v_2, v_3\}$ an embedding of $\{a, b, e\}$ with two possible Mappings:
  - Φ_1: a to v_2, b to v_1, e to v_3 .
  - Φ_2: a to v_2, b to v_3, e to v_3.

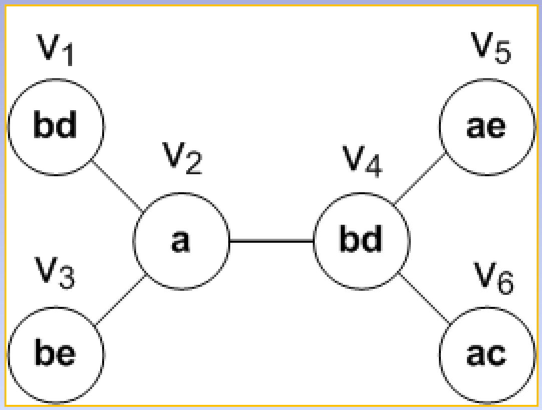


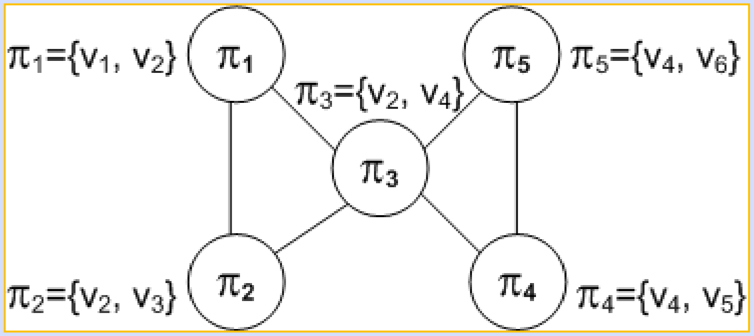
- $f(\pi)$ measures how tightly the mapped labels in the embedding π are connected. i.e., the inverse of diameter of π

- **SUPPORT:** Find all embeddings $\pi_1, \pi_2, \ldots, \pi_m$ of an itemset *I*. Define $\boldsymbol{sup(I) = \Sigma_i f(\pi_i)}$.

Neighborhood Association Model

- Overlap + Not Downward Closure !!!
- Use **maximum independent set** of all embeddings of an itemset. *(S. N. Bringmann, PAKDD'08)*
- $\boldsymbol{Sup(\mathrm{a, b}) = f(\pi_1) + f(\pi_4)}$.
- Downward Closure.
- Finding the maximum independent set is NP-hard

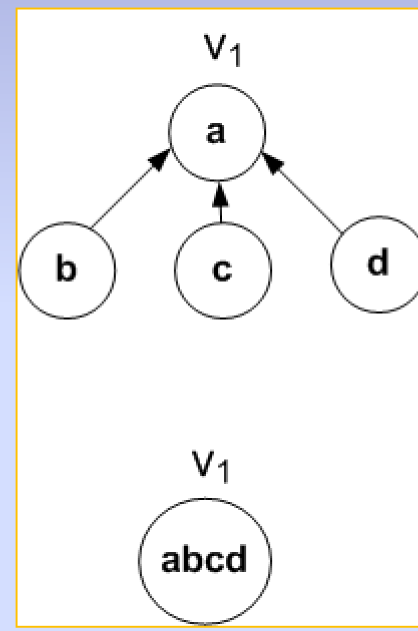




Embeddings of {a, b}

Information Propagation Model

- Influence Based Information Propagation.
- Information Propagation is modeled using First Order Markov Model.
- Labels are propagated with certain probability from each node to its neighbors.
- Labels are propagated independent to each other.



Information Propagation Model

- **NEAREST PROBABILISTIC ASSOCIATION (NPA):**
- If label l present in node u, $\boldsymbol{A_u(l) = 1}$.
- Otherwise, propagate l to u from its immediate neighbor v.
- $\boldsymbol{A_u(l) = A_v(l) \cdot e^{-\alpha}}$
- $\alpha > 0$ is the decay constant.
- Recursive to propagate beyond one hop.
- **SUPPORT:**

$$\boldsymbol{sup(\mathrm{I}) = (1 / |V|) \sum_{u \in V} A_u(l_1) \ldots A_u(l_m)}$$

$\mathbf{I} = \{l_1, \ldots, l_m\}$.

Information Propagation Model

- Downward Closure.
- Consistent with graph structure.

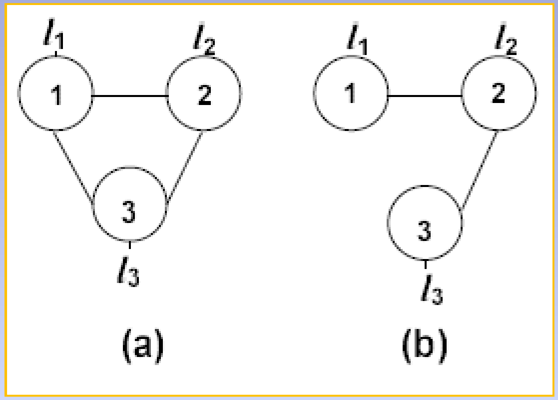
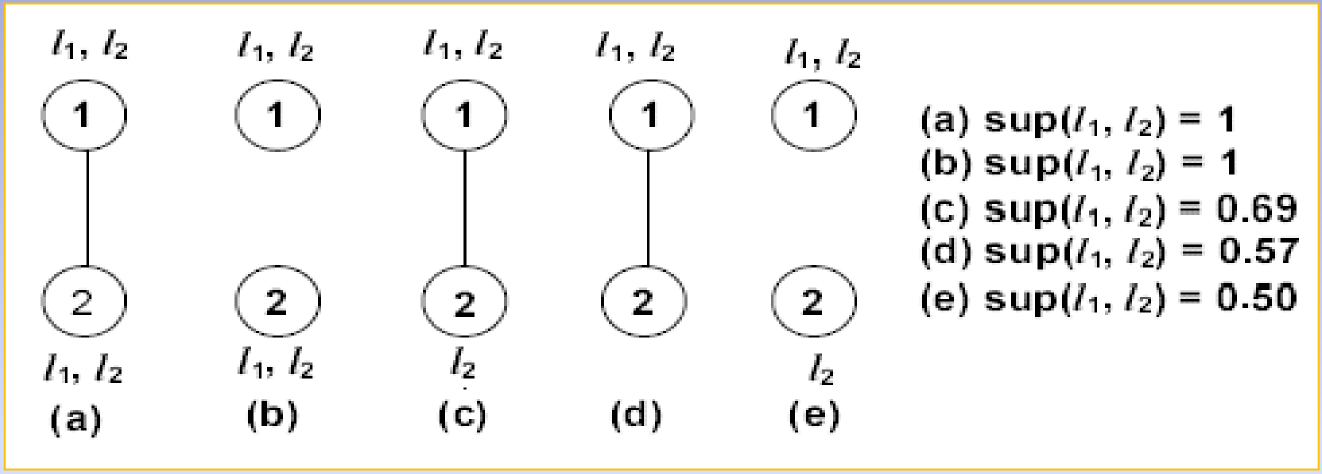


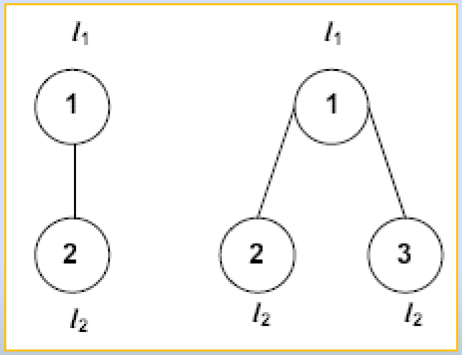
Table (a)			
	l_1	l_2	l_3
$node_1$	1	0.37	0.37
$node_2$	0.37	1	0.37
$node_3$	0.37	0.37	1
$\text{Sup}(l_1, l_2, l_3) = 0.14$			

Table (b)			
	l_1	l_2	l_3
$node_1$	1	0.37	0.14
$node_2$	0.37	1	0.37
$node_3$	0.14	0.37	1
$\text{Sup}(l_1, l_2, l_3) = 0.08$			

Information Propagation Model



- **PROBLEM WITH NEAREST PROBABILISTIC ASSOCIATION (NPA):**



$sup(l_1, l_2) = 0.37$!!!

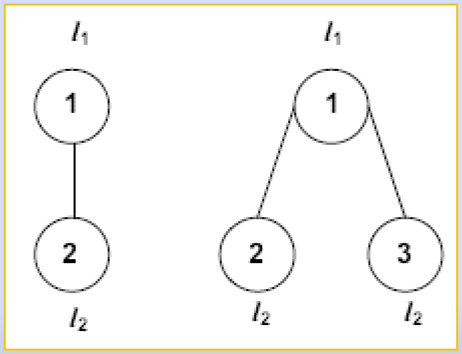
Information Propagation Model

- NORMALIZED PROBABILISTIC ASSOCIATION (**NmPA**):

$$A_u(l) = A_v(l) \,.\, [m/(n+1)]\, e^{-\alpha}$$

m = # of 1-hop neighbors of *u* containing label *l*.

n = # of 1-hop neighbors of *u*.

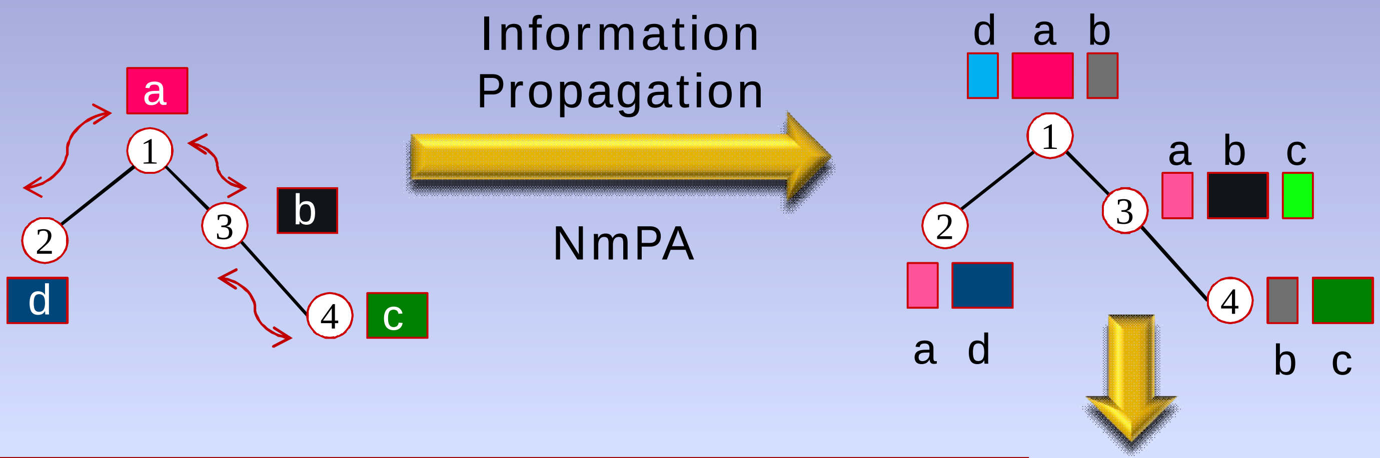


$$sup(l_1, l_2) = 0.37 \times (1/2) = 0.19$$
$$sup(l_1, l_2) = 0.37 \times (2/3) = 0.25$$

Roadmap

- Problem Formulation
  - Problem Definition
  - Preliminaries
- Framework
  - Neighborhood Association Model
  - Information Propagation Model
- **Probabilistic Itemset Mining**
- Experimental Results
- Conclusion

Probabilistic Itemset Mining



- **Frequent-Pattern (FP) Tree** cannot handle fractional association values because of the new definition of Support.
- Modify FP Tree Structure and Algorithm.
- *C. C. Aggarwal et. al (KDD '09), Bernecker et. al (KDD '09).*

	a	b	c	d
1	1.00	0.12	0.00	0.12
2	0.19	0.00	0.00	1.00
3	0.12	1.00	0.12	0.00
4	0.00	0.19	1.00	0.00

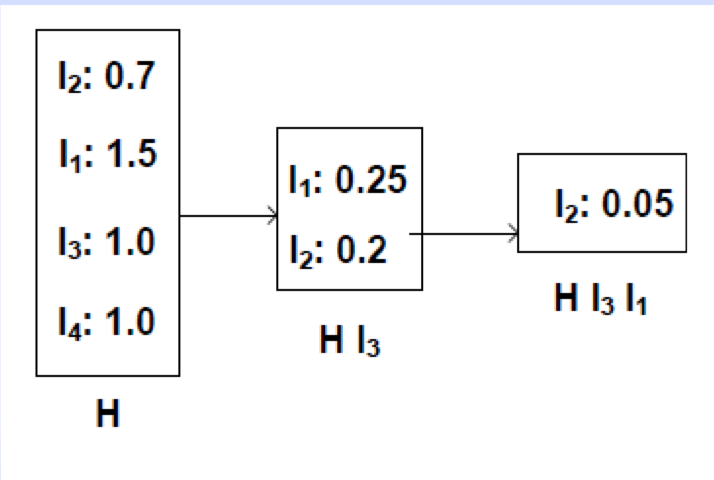
Frequent Itemset Mining (Probabilistic)

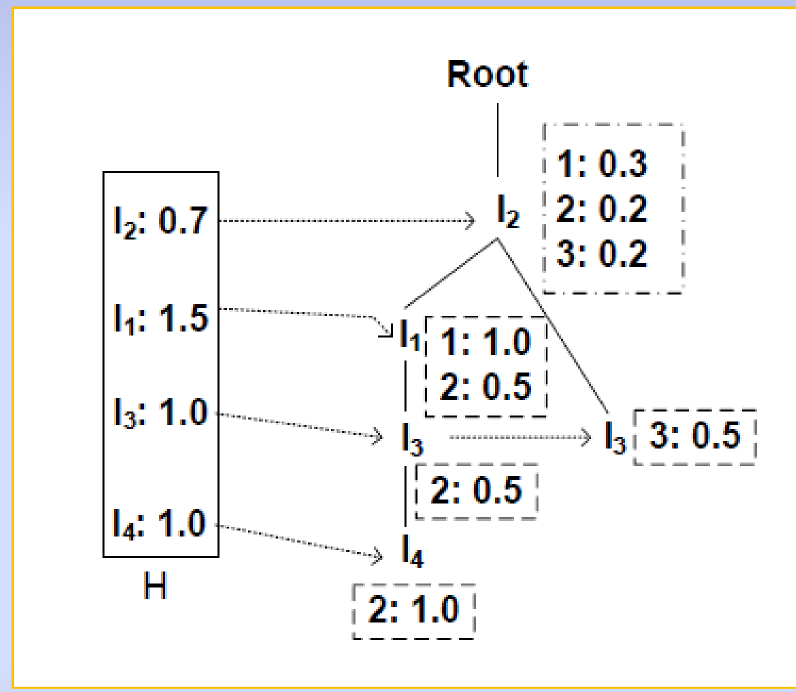
Probabilistic Itemset Mining

- **Probabilistic FP-Growth (pFP)**:

associating a **bucket** with each node of the FP-tree.

transaction id	l_1	l_2	l_3	l_4	l_5
1	1	0.3	0	0	0.1
2	0.5	0.2	0.5	1	0
3	0	0.2	0.5	0	0.05





Probabilistic Itemset Mining

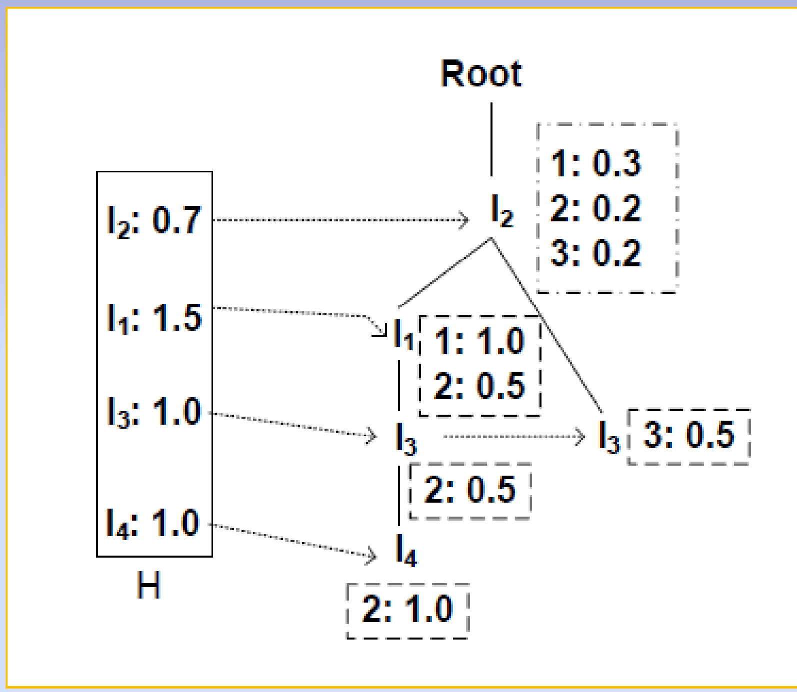
- PROBLEMS WITH PROBABILISTIC FP-TREE (pFP): slow because of frequent disk access to load and store the buckets.

- Is it possible to approximate the buckets so that the complete tree can be loaded in the main memory?

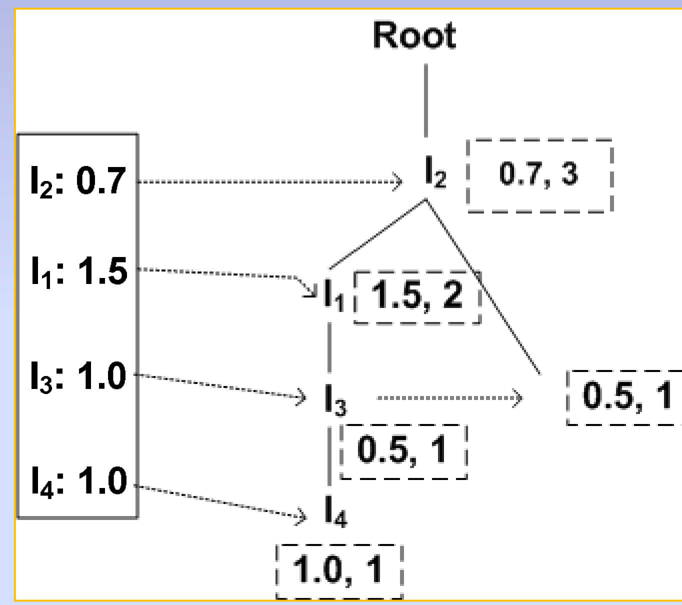
- **Approximate FP-Tree (aFP)**

Probabilistic Itemset Mining

- APPROXIMATE FP-TREE (aFP):



$sup(l_1, l_2) = 0.4$



$sup(l_1, l_2) = 0.35$

$$\widetilde{A}(l_x, l_y) = \frac{sum(v_x) \cdot sum(v_y)}{max\{occurrence(v_x), occurrence(v_y)\}}$$

Top-k Interesting Pattern Mining

- How to measure "Interesting-ness"? – **Randomization Test**.
- Generate graph *Q* from graph *G* by randomly swapping the labels among nodes. Let, *p* and *q* be the support values of itemset *I* in *G* and *Q* respectively. High difference indicates interestingness.
- G-test Score: $$p \cdot \ln \frac{p}{q} + (1-p) \cdot \ln \frac{1-p}{1-q}$$
- Vertical Pruning by *Yan et. al (SIGMOD '08).*
- Proximity Patterns minus Frequent Patterns.

Roadmap

- Problem Formulation
  - Problem Definition
  - Preliminaries
- Framework
  - Neighborhood Association Model
  - Information Propagation Model
- Probabilistic Itemset Mining
- **Experimental Results**
- Conclusion

Experimental Results

- DATASET:

	# of Nodes	# of Edges	# of Labels	Avg. # of Labels/ Node
Last.FM	6,899	58,179	6,340	3
Intrusion	200,858	703,020	1,000	25
DBLP	684,911	7,764,604	130	9

- EFFICIENCY:

	Last.FM	Intrusion	DBLP
NmPA	2.0 sec	5.0 sec	187.0 sec
FP-Tree Formation	1.0 sec	10.0 sec	89.0 sec
Top-k Mining	4.0 sec	2.0 sec	254.0 sec

Experimental Results

❑ EFFECTIVENESS (Last.FM):

Proximity Patterns

#	Proximity Patterns	Score
1	Tiësto, Armin van Buuren , ATB	0.62
2	Katy Perry, Lady Gaga, Britney Spears	0.58
3	Ferry Corsten, Tiësto, Paul van Dyk	0.55
4	Neaera, Caliban, Cannibal Corpse	0.52
5	Lacuna Coil, Nightwish, Within Temptation	0.47

- ATB, Paul van Dyk – **German DJ**
- Tiesto, Ferry Corsten, Armin van Buuren – **Dutch DJ**
- Britney Spears, Lady Gaga, Katy Gaga – **American Female Pop Singers**
- Neaera, Caliban, Cannibal Corpse – **Death Metal Bands**
- Lucuna Coil, Nightwish, Within Temptation – **Gothic Metal Bands**

Experimental Results

- EFFECTIVENESS (Intrusion):

#	Interesting Patterns	Score
1	Ping_Sweep, Smurf_Attack	2.42
2	TFTP_Put, Audit_TFTP_Get_Filename, ICMP_Flood, Ping_Flood	2.32
3	TCP_Service_Sweep, Email_Error	1.21
4	HTML_Outlook_MailTo_Code_Execution, HTML_NullChar_Evasion	1.15
5	SQL_SSRP_Slammer_Worm, SQL_SSRP_StackBo	0.88

Proximity Patterns

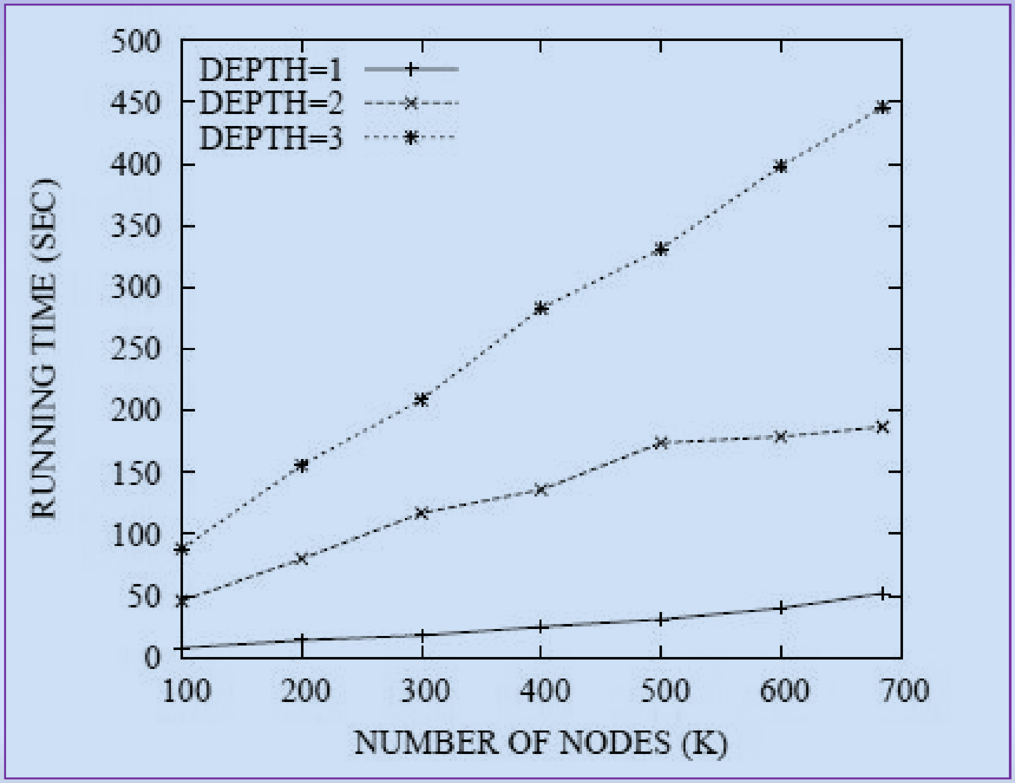
#	Interesting Patterns	Score
1	ICMP_Flood, Ping_Flood	0.94
2	Email_Error, SMTP_Relay _Not_Allowed, HTML_Null Char_Evasion	0.94
3	Image_RIFF_Malformed, HTML_NullChar_Evasion	0.90
4	TFTP_Put, Ping_Flood, Audit_TFTP_Get_Filename	0.80
5	Email_Command_Overflow, Email_Virus_Double_Extension, Email_Error	0.75

Proximity Patterns Minus Frequent Patterns

Experimental Results

❑ SCALIBILITY

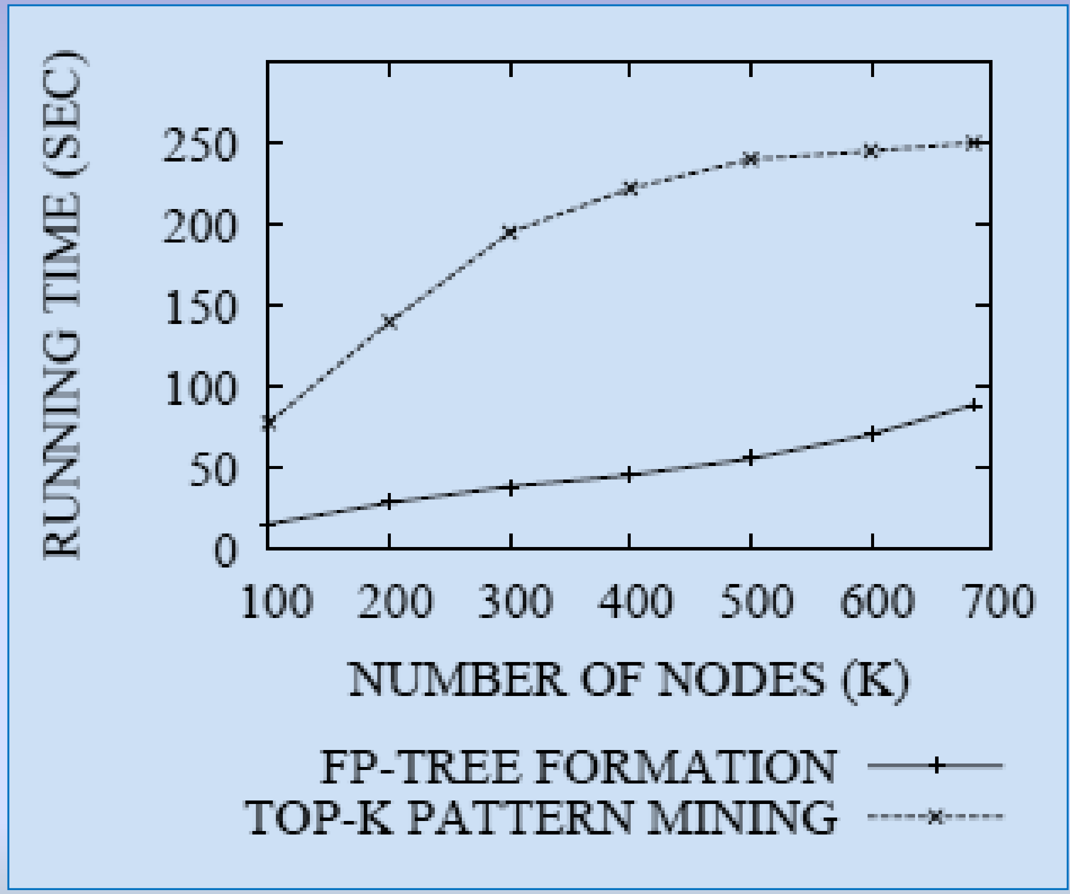
Information Propagation (NmPA) Time vs. No. of Nodes



Experimental Results

- SCALIBILITY

Mining Time
vs.
No. of Nodes



Experimental Results

- pFP (Exact Mining) vs. aFP (Approximate Mining) [Last.FM]:

#	Proximity Patterns	Score
1	Tiësto, Armin van Buuren , ATB	0.62
2	Katy Perry, Lady Gaga, Britney Spears	0.58
3	Ferry Corsten, Tiësto, Paul van Dyk	0.55
4	Neaera, Caliban, Cannibal Corpse	0.52
5	Lacuna Coil, Nightwish, Within Temptation	0.47

aFP (Approximate Mining)

#	Proximity Patterns	Score
1	Katy Perry, Lady Gaga, Britney Spears	0.58
2	Ferry Corsten, Tiësto, Paul van Dyk	0.55
3	Tiësto, Armin van Buuren, ATB	0.55
4	Neaera, Caliban, Cannibal Corpse	0.51
5	Lacuna Coil, Nightwish, Within Temptation	0.46

pFP (Exact Mining)

Steps	aFP(approximate)	pFP(exact)
FP-tree Formation	1.0	3.0
Top-k Pattern Mining	4.0	21.0

Table 10: Runtime Comparison (sec) (Last.fm)

Roadmap

- Problem Formulation
  - Problem Definition
  - Preliminaries
- Framework
  - Neighborhood Association Model
  - Information Propagation Model
- Probabilistic Itemset Mining
- Experimental Results
- **Conclusion**

Conclusion

- Novel Concept of Proximity Pattern Mining in Large Graphs.
- Neighborhood Association Model and Information Propagation Model. Probabilistic Itemset Mining Algorithms.
- Effective, Efficient and Scalable framework.
- How to determine the optimal propagation measure and depth?

Questions ??

Thank You !