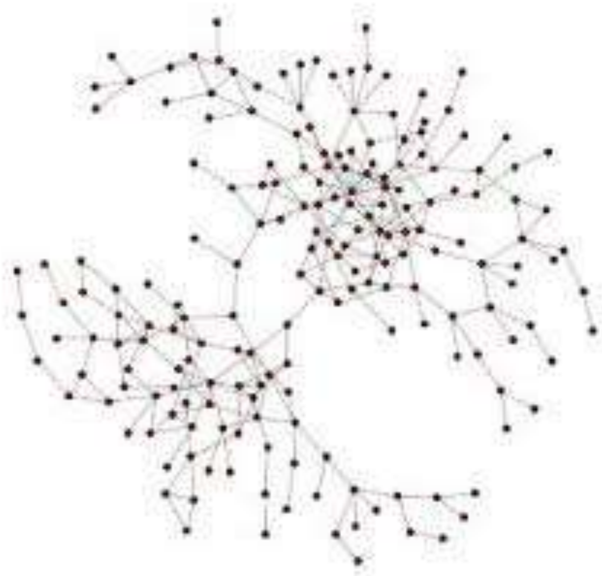
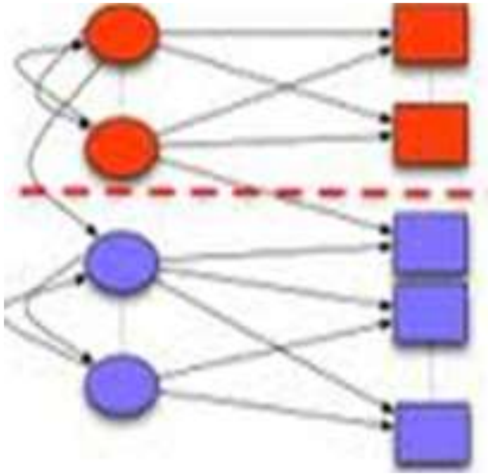
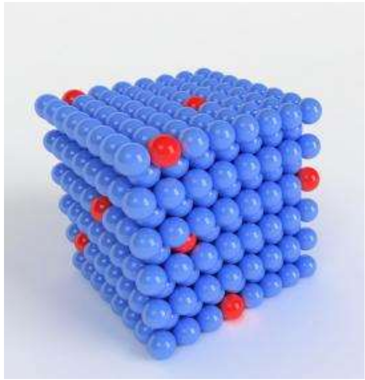
Relevance Search and Anomaly Detection using Bipartite Graphs



Student: Vikramaditya Jakkula

Committee: Dr. Larry Holder (chair),
Dr. Ananth Kalyanaraman, Dr. K.C. Wang





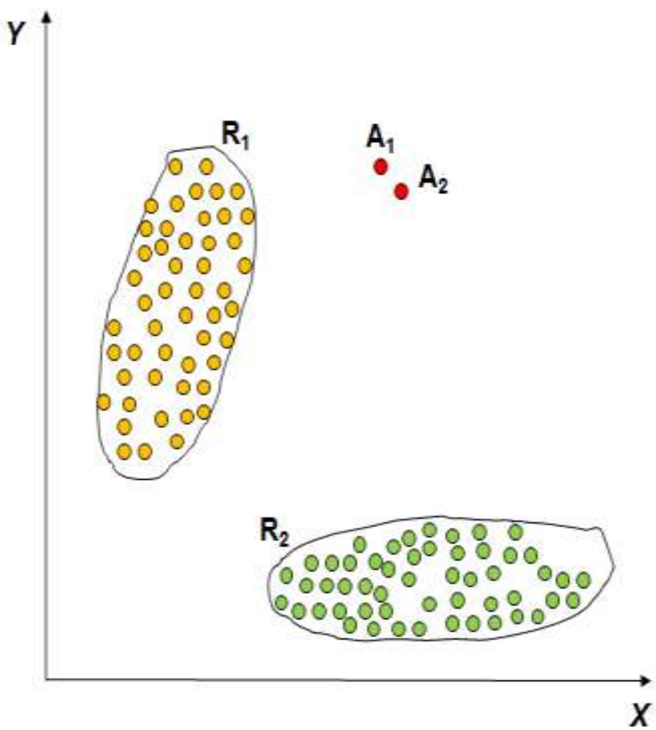
Washington State University
World Class. Face to Face.

What is an Anomaly?

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data.

- Anomaly pattern is a pattern in the data that does not conform to the normal behaviour.

- Also referred to as outliers, exceptions, peculiarities, surprise, etc.



Anomaly Detection

- Normal Pattern ➔ Deviations ➔ Anomaly Pattern
- Anomaly detection deals with deviations (data points) that are very different from the "normal" activities (rest of the data points)
- General Steps
  - Identify "normal" data
  - Define "anomalous" data
  - Construct useful set of features
  - Use outlier/anomaly detection algorithm
    - Statistics based
    - Distance based
    - Model based

Challenges in Anomaly Detection

- Defining normal behavior from abnormal
- Availability of ground truth
- Dealing with Noise
- Normal behavior evolves with time

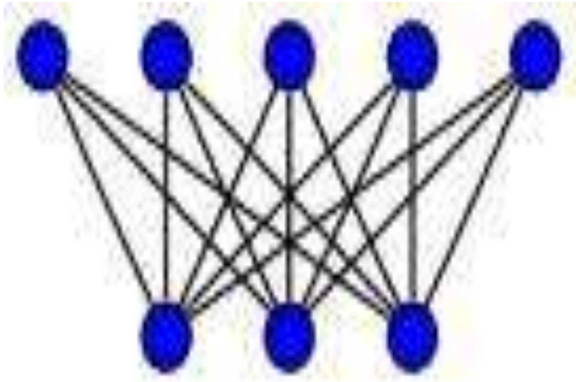
Application of Anomaly Detection

- Fraud detection (credit card, phone, etc)
- Spam detection
- Image Analysis
  - detecting geographic hotspots
- Network intrusion detection
- Health Informatics – Epidemic Outbreaks, and more.

Introduction

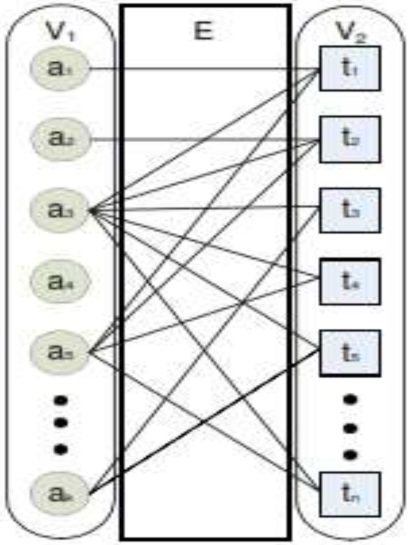
- Bipartite Graph:

A graph whose vertices can be divided into two disjoint sets *U* and *V* such that every edge connects a vertex in *U* to one in *V*; that is, *U* and *V* are independent sets.



- Relevance Search:

Given a node “a” in v1, the relevancy(via a relevancy score) of all other nodes in V1 to “a” forms relevance search.

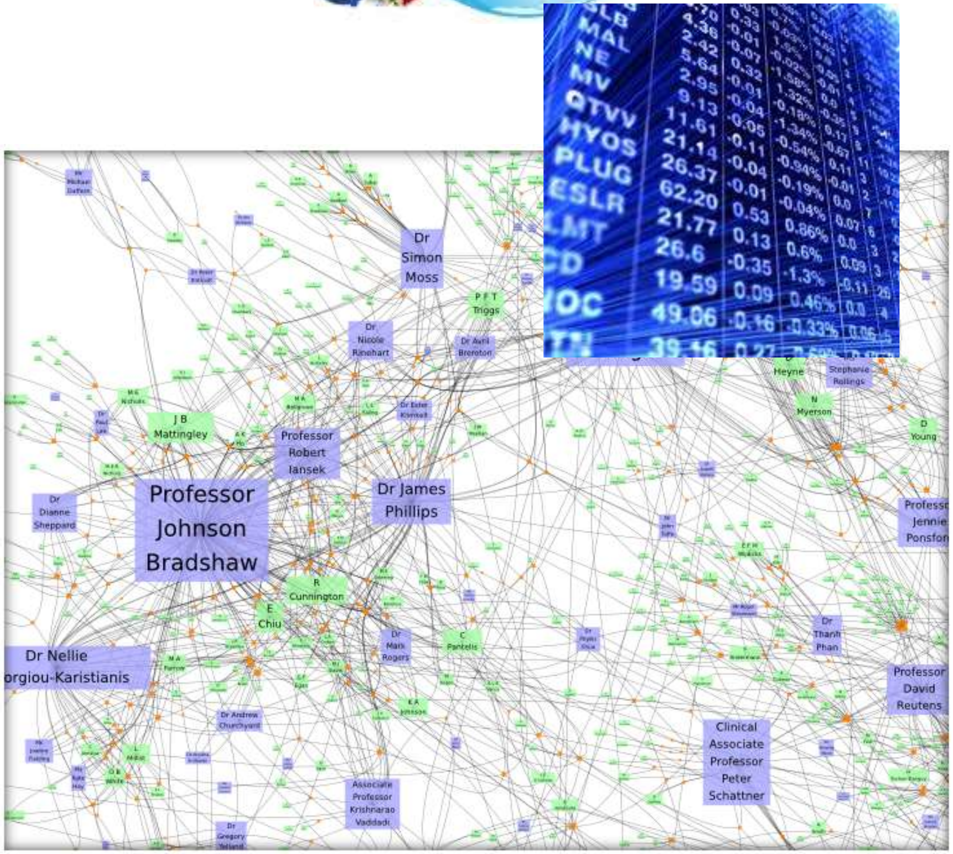


- Anomaly Detection:

Given a node a in V1, computing the normality score in V2 for all nodes connected to a and ones with lowest score are deemed anomaly.

Application of Bipartite Graphs

- P2P System: Files and Peers form V1 and V2 and edge is download or upload action.
- Stock Markets: Traders and stocks form V1 and V2 and the buying and selling action form the edge.
- Research Publications: Conferences and authors form V1 and V2 and edge represents the publication



Related Work

- Graph Partitioning:
  - Spectral partitioning methods
  - Information theoretic approaches
  - And so forth.
  - No particular reason as why they picked METIS compared to others.
- Outlier Detection:
  - Finds outlier edges in a graphs
  - Finding anomalous substructures
  - Key trade off was computation cost.
- Random-Walks: Includes pagerank, similarity rank and other ranking approaches where random walk is used.
- Collaborative Filtering: Not applicable to anomaly detection

Data Representation

- Data is viewed as bipartite graph
- The graph G is conceptually stored in a k-by-n matrix M, where M(i, j) is the weight of the edge < i, j >

$$M_{k\times n} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & 1 & 1 & \dots & 1 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & 1 & 0 & \dots & 1 \\ \dots & & & & & & \\ 0 & 0 & 1 & 0 & 1 & \dots & 1 \end{pmatrix}$$

- The value can be 0/1 for an unweighted graph, or any nonnegative value for a weighted graph.

Proposed Method: Relevance Search

Algorithm RSE(Exact RS)

Input: node a, bipartite matrix M, restarting probability c, tolerant threshold ε

0. initialize $\vec{q}_a = 0$ except the a-th element is 1 ($q_a(a) = 1$)
1. construct M_A (see Equation 1) and $P_A = col\_norm(M_A)$
2. while ($|\Delta\vec{u}_a| > \varepsilon$)
$\vec{u}_a = (1-c)P_A\vec{u}_a + c\vec{q}_a$
3. return $\vec{u}_a(1:k)$

Algorithm RSA(Approximate RS)

Input: the bipartite graph G, the number of partitions κ, input node a

0. divide G into κ partitions $G_1 \ldots G_\kappa$ (one-time cost)
1. find the partition G_i containing a
2. construct the approximate bipartite matrix M' of G_i (ignore the edges cross two partitions)
3. apply RS_E on a and M'
4. set 0 relevance scores for the nodes that are not in G_i

- Exact RS has slow convergence rate
- Approximate RS uses graph partition (METIS is used) to result in k non-overlapping sub graphs of about same size.

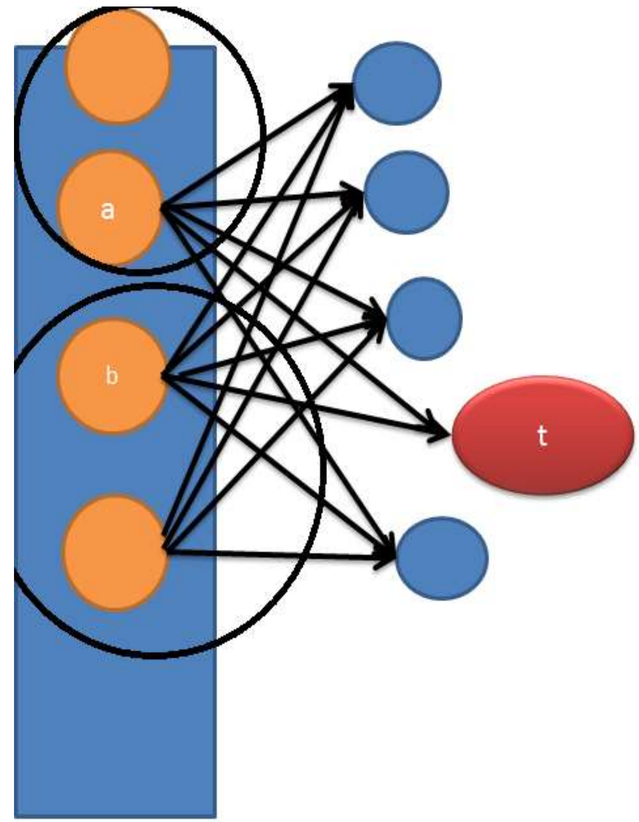
Proposed Method: Anomaly Detection

- The normality score ns(t) can be any function.

Algorithm AD(Anomaly Detection)

Input: input node t, bipartite transition matrix P

0. find the set $S_t = \{a_1, a_2, ...\}$ such that $\forall a_i \in S_t, < a_i, t > \in E$.
1. compute all the relevance score vectors $\vec{R}$ of $a \in S_t$
2. construct the similarity matrix RS_t from $\vec{R}$ over S_t
3. apply the score function over RS_t to obtain the final normality score $ns(t)$
4. return $ns(t)$



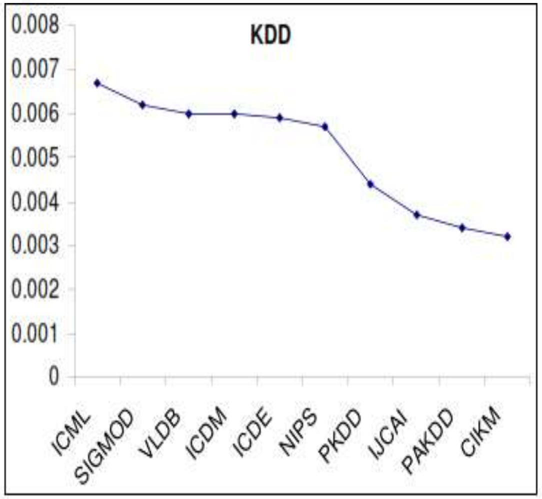
Experiment Settings

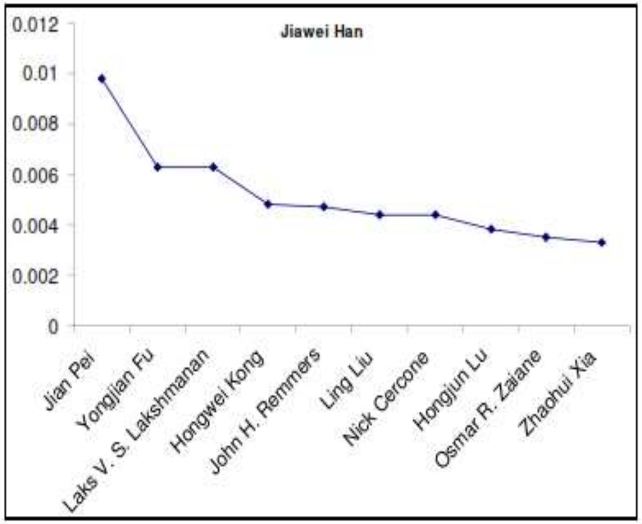
- Three real world datasets were used.
- Conference –Author:
  - Row represents conferences and column an author.
  - On an average, every conf. has 510 authors, every author publishes in 5 conf.
- Author – Paper:
  - Row represents author, and column represents paper.
  - On an average every author has 3 papers, every paper has 2 authors.
- IMDB:
  - Row represents actor/actress and column represents movies.
  - On an average every actor/actress plays in 4 movies, and every movie has 11 actors/actress.

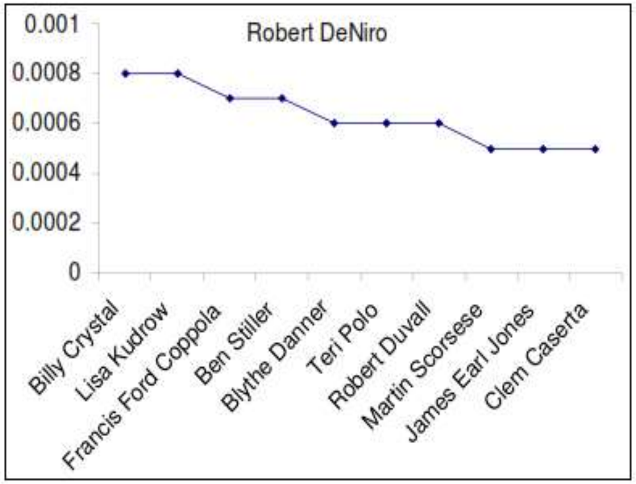
Dataset	Rows	Columns	Nonzeros	Weighted
CA	288590	2687	661535	yes
AP	315688	471514	1073168	no
IMDB	553388	204000	2269811	no

Evaluation of Exact RS

- Goal: To check whether nodes with high relevance scores are closely related to query node.
- CA dataset:
  - Displays top 10 conferences for a selected example of KDD.
- AP Dataset:
  - Picked an arbitrary example and found close collaborators.
- IMDB Dataset:
  - Just picked Robert De Niro
  - Issue to note is sequel of movies

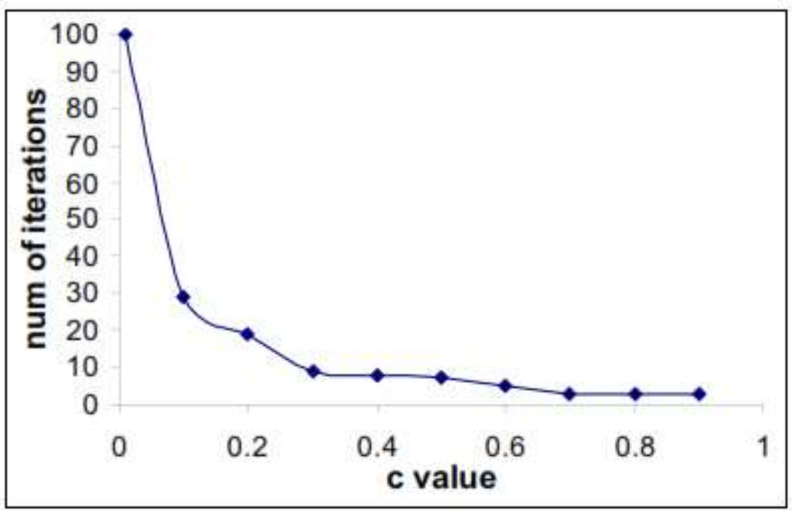


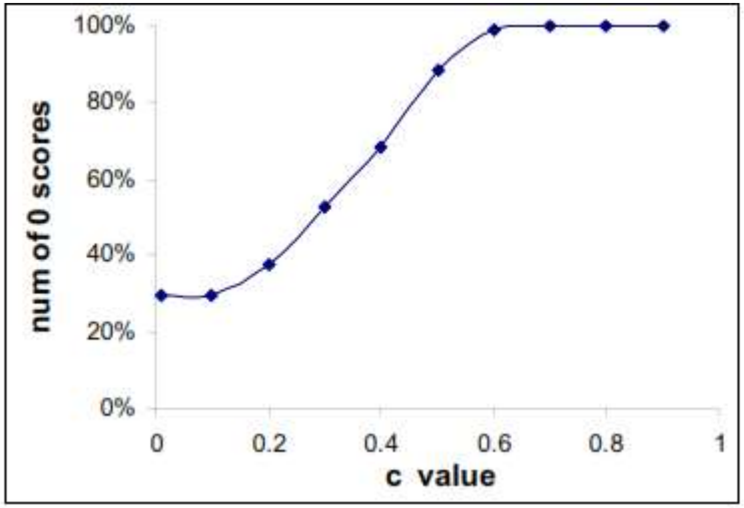




Evaluation of Convergence

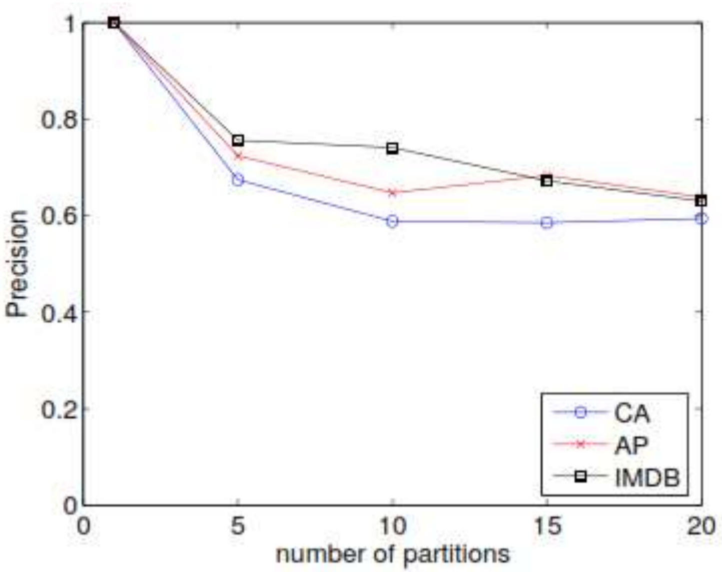
- Goal: Evaluate the variation of restart probability(c) and convergence threshold(E) to get the best values
- E has lesser effect on relevance score and is chosen to be 0.1
- No of iterations drop when C is larger which means the method converges quickly.
- For efficiency and effectiveness the value of C = 0.15 and e = 0.1

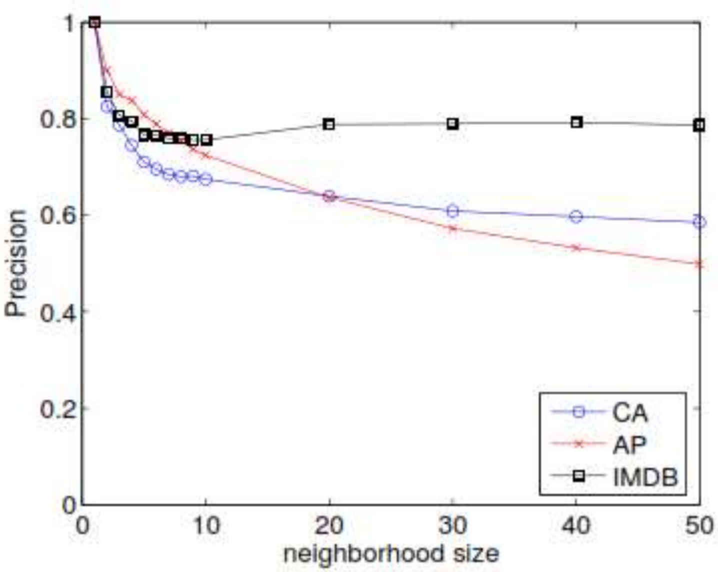




Evaluation of Approximate RS

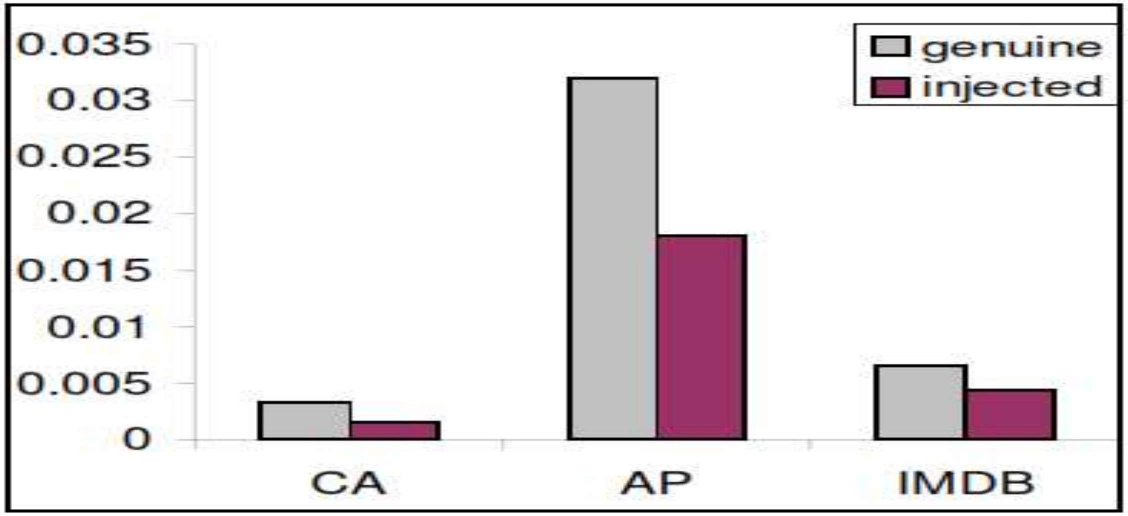
- We observe with graph partition leads to reduced precision in all three datasets.





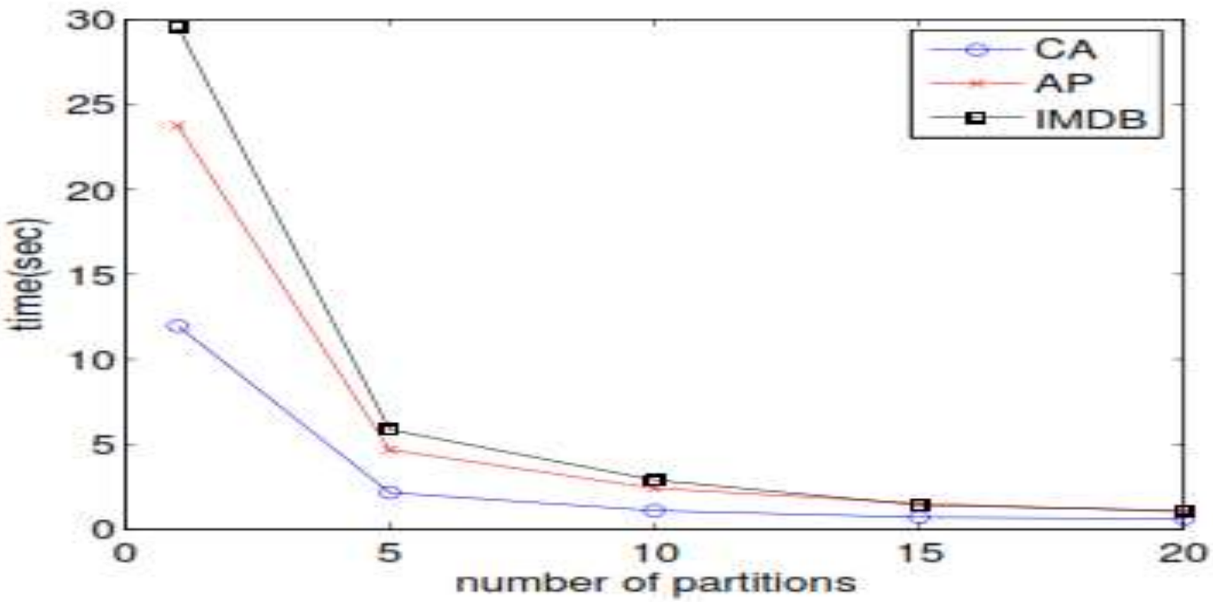
Evaluation of Anomaly Detection

- Artificial anomalies injected (100 injected)
- We note that the injected anomalies can be easily identified as the normality scores are low in all three datasets.



Evaluation of Computational Cost

- All computation comes down to RS execution
- An approach using partitions is computationally effective.



Anomaly Detection in Smart Environments

- Standardization of Smart Sensor Datasets
- Apply to activity abnormality detection (also power consumption analysis)
- Feedback to Learning Models
- Reminder systems and prompting system performance improvement
- Evaluation of human lifestyles & improvement suggestions
- Power Anomaly analysis

Bipartite Graphs based approach to Anomaly Detection in Smart Environments

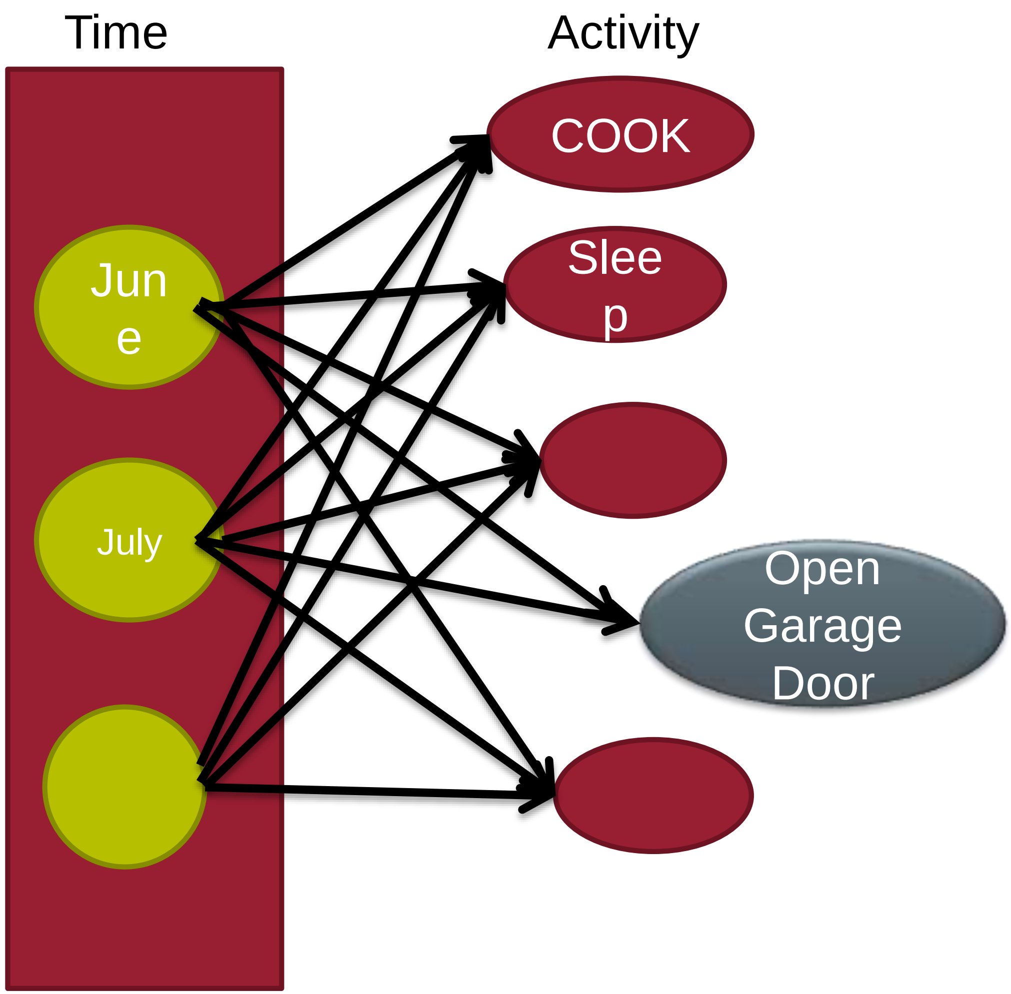
- We use three Scenarios/tasks for illustrations:

Activity Abnormality Detection

Recommendation/Prompting System

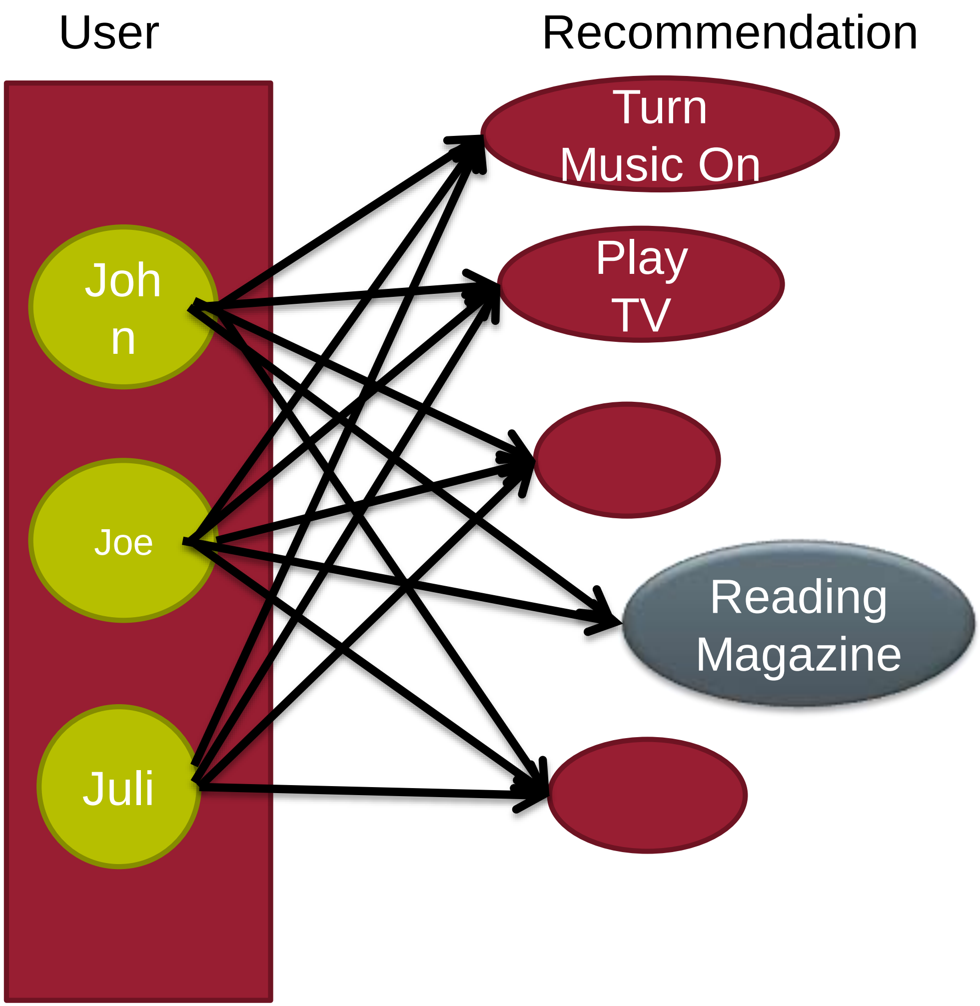
Power Abnormality Detection

Activity Abnormality Detection



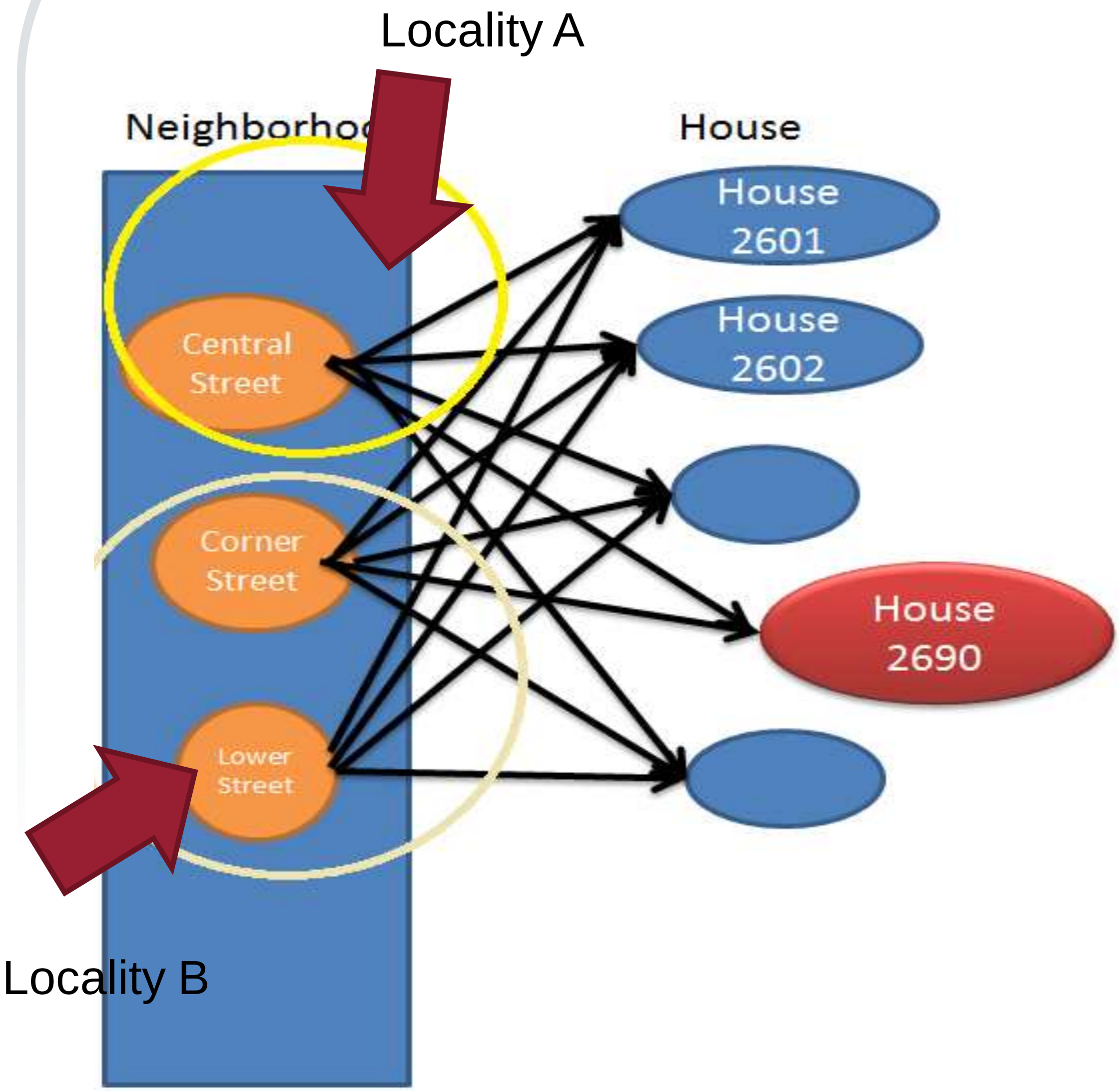
- Time and Activities form the two set of vertices and the edges are the frequency of occurrence.
- Help find anomalous activities which can be used to investigate, automate or inputted as feedback to learning algorithms.

Recommendation /Prompting System with Anomaly analysis



- User Profile and recommendation are two set of vertices and the no of times this was suggested is the edge.
- Helps identify lifestyle changes and abnormalities which can also help evaluate emotional state of the user.

Power Abnormality Detection



- Neighborhood and house are vertices, with power consumption as edge.
- Help find abnormal power usage and help residents make better decisions in power consumption.

Conclusion

- Authors address relevance search and anomaly detection on bipartite graphs with solution:
    - Fast convergence
    - Scalable
    - Simplistic to implement
    - Easy to interpret results

- Anomaly detection is very niche problem in smart environment world and has wide benefits in this domain, with large application opportunities.

Reference: (1) All general or generic images are from www.google.com
(2) Jimeng Sun, Huiming Qu, Deepayan Chakrabarti, and Christos Faloutsos (2005).
Relevance search and anomaly detection in bipartite graphs. SIGKDD Explorations Newsletter, 7(2):48-55.