SEEKING STABLE CLUSTERS IN THE BLOGOSPHERE

VLDB 2007, VIENNA

Nilesh Bansal, Fei Chiang, Nick Koudas

University of Toronto

Frank Wm. Tompa

University of Waterloo

The Blogosphere

UNIVERSITY OF TORONTO

- The new way to communicate
  - Millions of text articles posted daily
  - From all over the globe
  - A wide variety of topics, from sports to politics
  - Forms a huge repository of human generated content
- A high volume temporally ordered stream of text documents
- Challenge: discover persistent chatter

BlogScope

UNIVERSITY OF TORONTO

- Live blog search and analysis engine
  - Tracking over 13 million blogs, 100 million posts
  - Serves thousands of daily visitors
- Visit: **www.blogscope.net**

Demo Today: 4:30 - 6:00 pm

Nilesh Bansal, Nick Koudas, BlogScope: A System for Online Analysis of High Volume Text Streams, VLDB *2007*, Demonstration Proposal

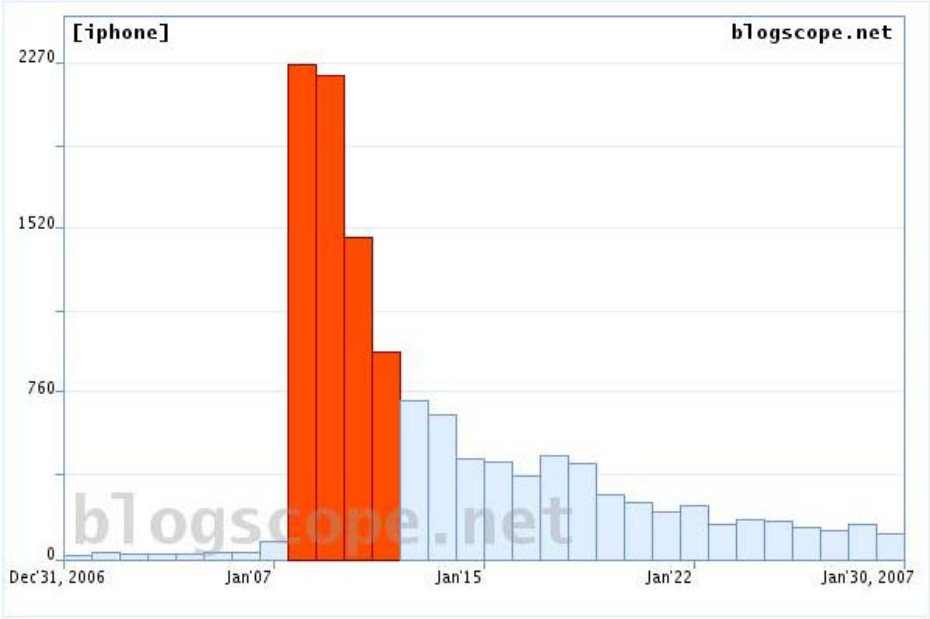
Nilesh Bansal, Nick Koudas, Searching the Blogosphere, *WebDB 2007*

UNIVERSITY OF TORONTO

Persistent Chatter

UNIVERSITY OF TORONTO

- Apple iPhone – January 2007
  - Jan first week: Anticipation of iPhone release
  - Jan 9th: iPhone release at Macworld
  - Jan 10th: Lawsuit by Cisco
  - Jan third week: Decrease in chatter about iPhone



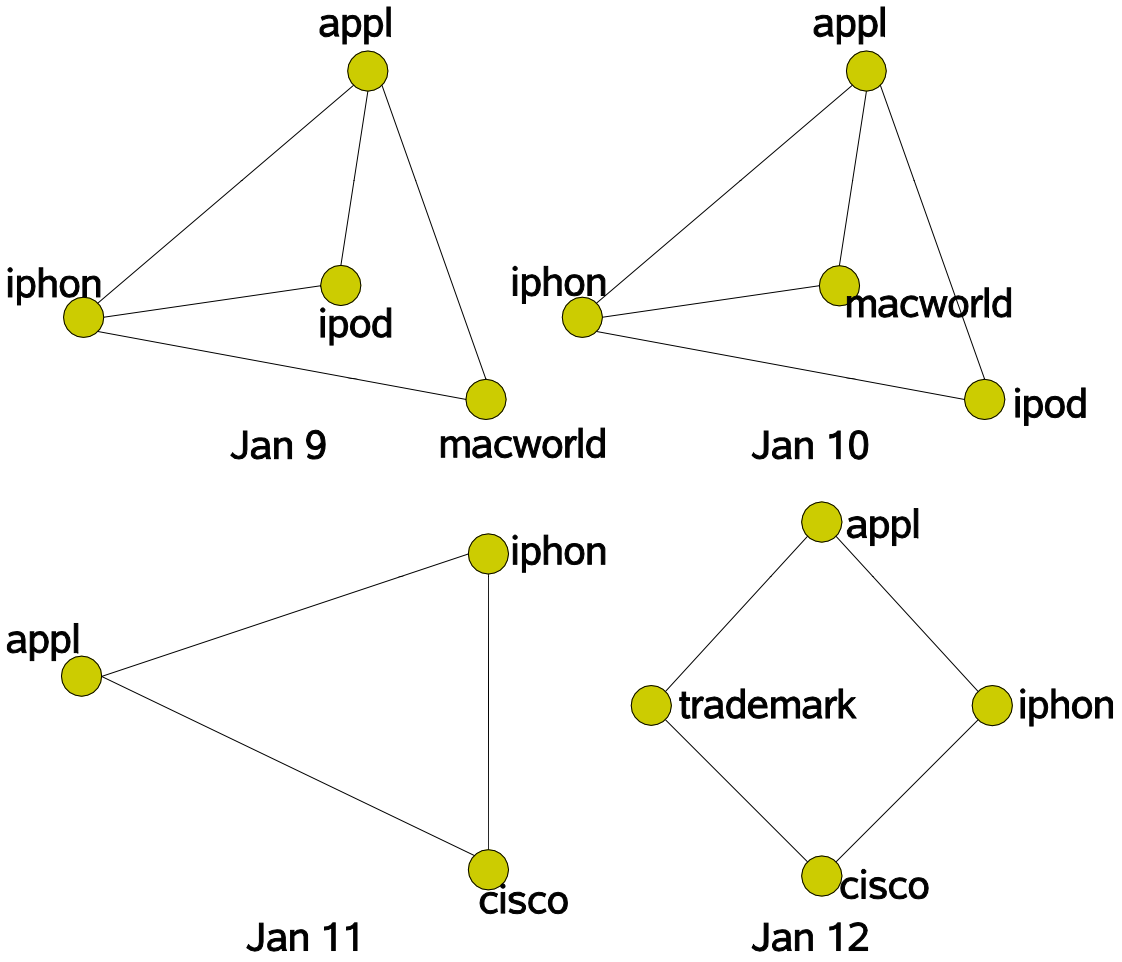
Keyword Clusters

UNIVERSITY OF TORONTO

- When there is a lot of discussion on a topic, a set of keywords will become correlated
  - Elements in this keyword set will frequently appear together
  - These keywords form a cluster
- Keyword clusters are transient
  - Associated with time interval
  - As topics recede, these clusters will dissolve

Stable Clusters - Apple iPhone

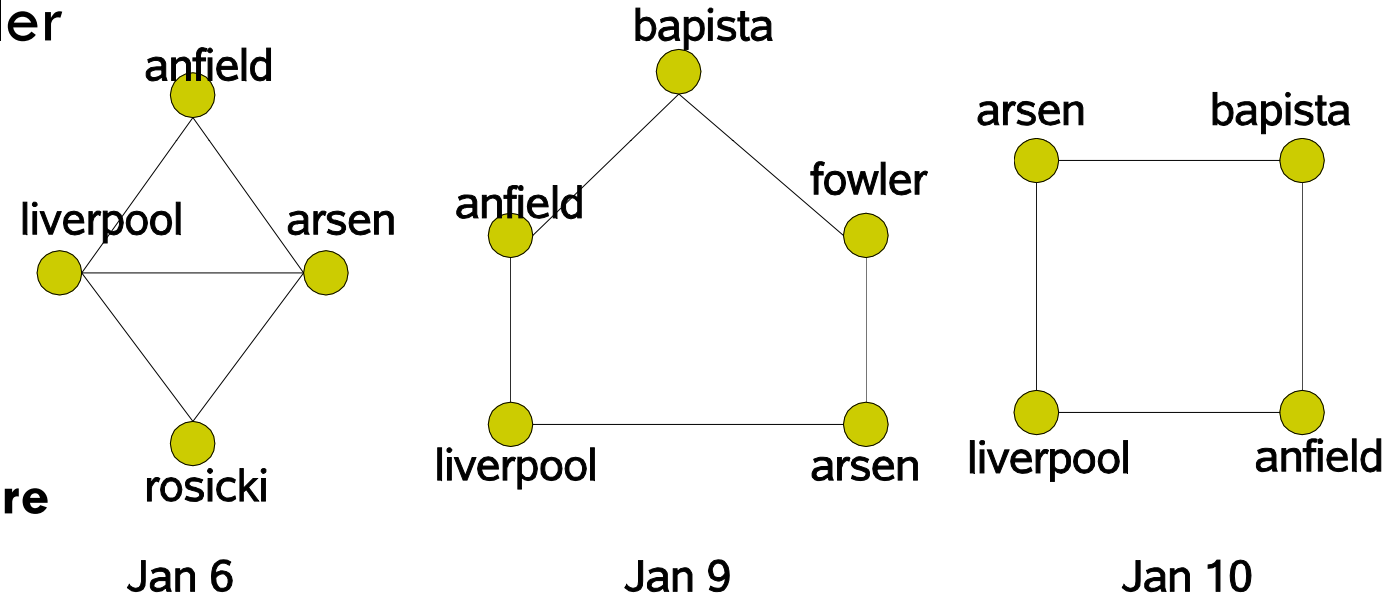
- Persistent for 4 days
- Topic drifts
  - Starts with discussion about Apple in general
  - Moves towards the Cisco lawsuit



Note: All keywords are stemmed

Gap in Clusters

- Three clusters are shown for Jan 6, 9 and 10 2007; no clusters were discovered for Jan 7 and 8 (related to this topic)
- English FA cup soccer game between Liverpool and Arsenal with double goal by Rosicky at Anfield on Jan 6. The same two teams played again on Jan 9,with goals by Bapista and Fowler



Note: keywords are stemmed

Why Stable Clusters

UNIVERSITY OF TORONTO

- Information Discovery
  - Monitor the buzz in the Blogosphere
  - "What were bloggers talking about in April last year?"
- Query refinement and expansion
  - If the query keyword belongs to one of the cluster
- Visualization?
  - Show keyword clusters directly to the user
  - Or show matching blogs

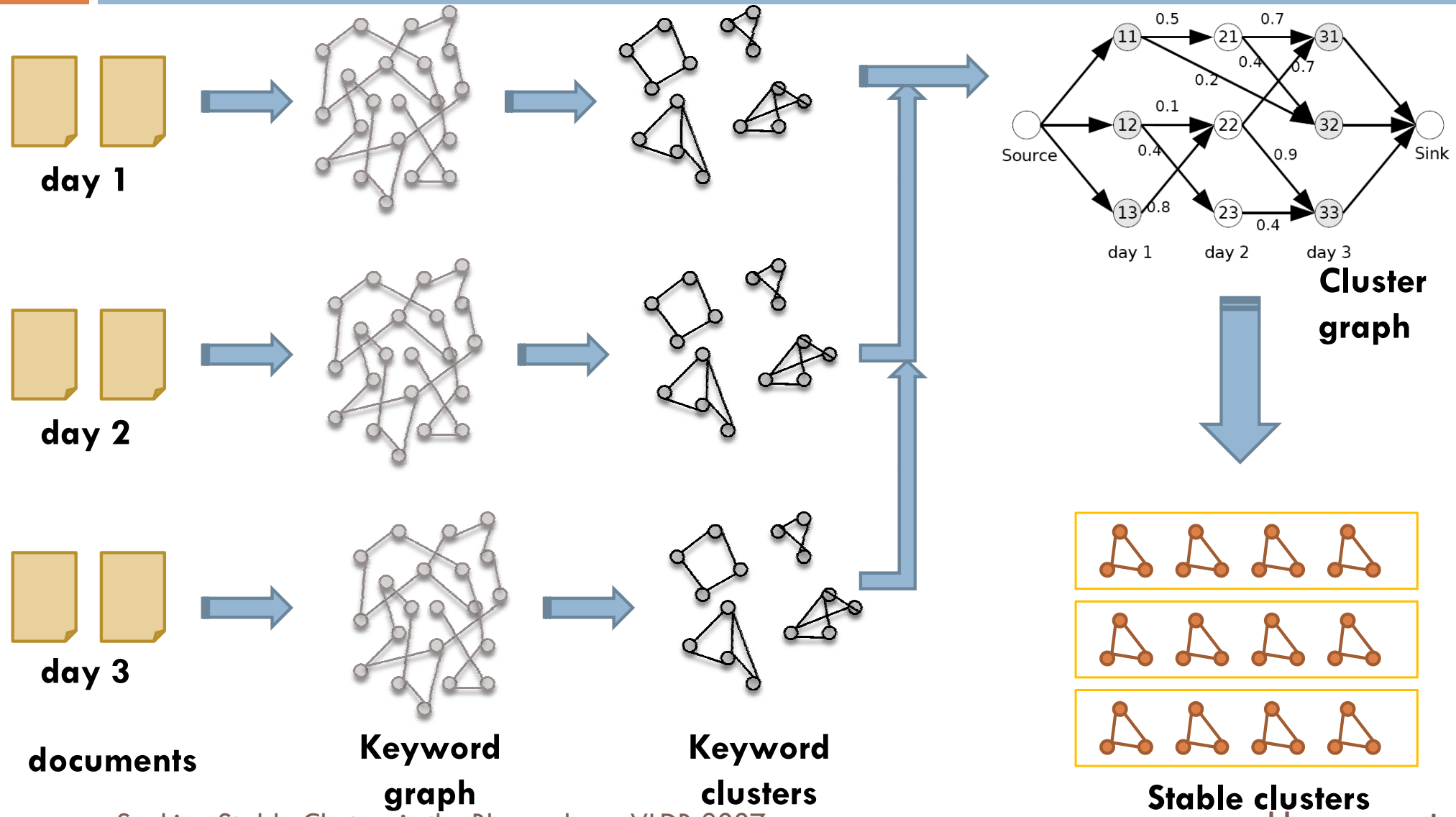
Overview

UNIVERSITY OF TORONTO

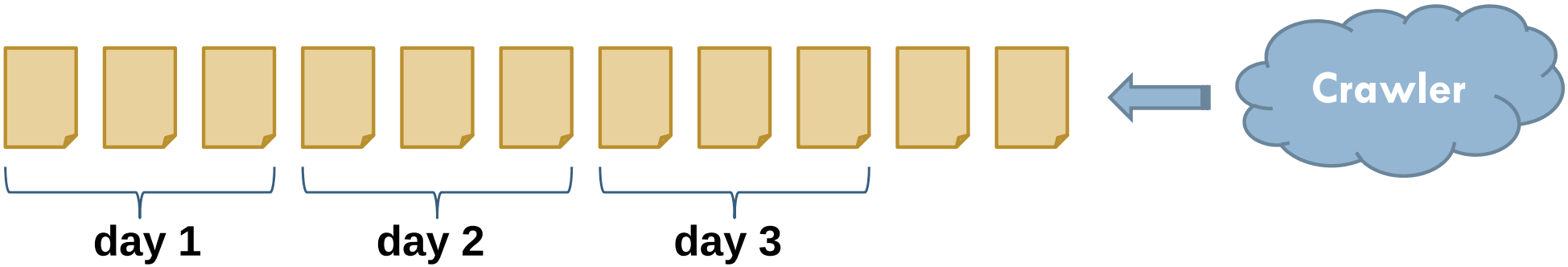
- Efficient algorithm to identify keyword clusters
  - BlogScope data contains over 13M unique keywords
  - Applicable to other streaming text sources
    - Flickr tags, News articles
- Formalize the notion of stable clusters
- Efficient algorithms to identify stable clusters
  - BFS, DFS and TA
  - Amenable to online computation over streaming data
- Experimental evaluation

Pipeline

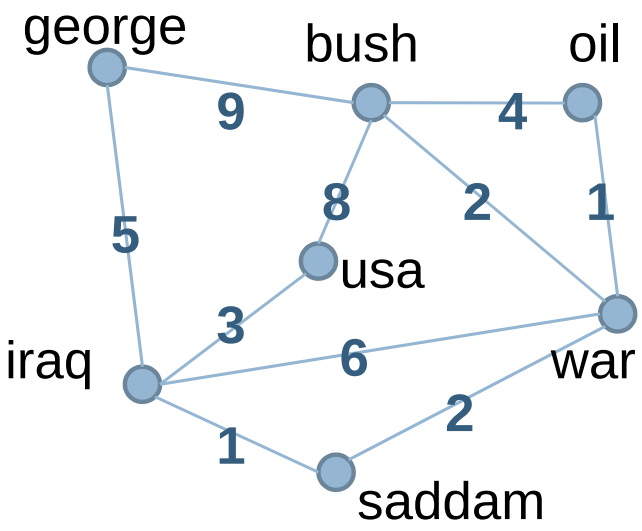
UNIVERSITY OF TORONTO



Keyword Graph



- One undirected graph for each day
  - Each keyword forms a node
  - Edge weight = number of documents in which both the keywords occur



Graph for *i*th day

Pruning the Graph

UNIVERSITY OF TORONTO

- Keep only strong keyword associations
- Assess two way association between keyword pairs [Manning & Schutze, 1999]
  - Pearson Chi-square test
  - Correlation coefficient

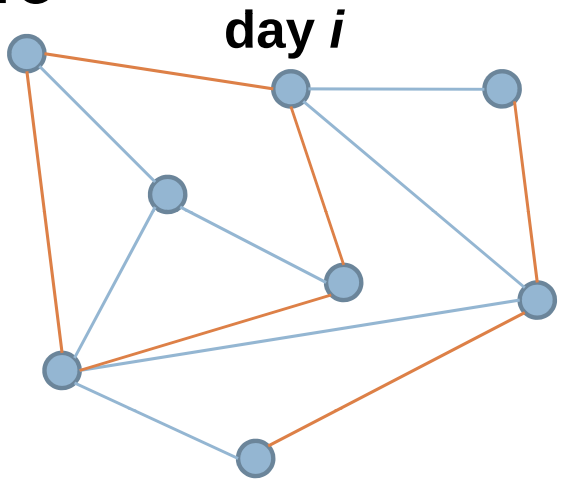
Date	File Size	# keywords	# edges
Jan 6 2007	3027MB	2.8 million	138 million
Jan 7 2007	2968MB	2.8 million	135 million

Keyword graph – after stemming, and removing stop words

Chi-square and Correlation

UNIVERSITY OF TORONTO

- Perform a single pass on the graph
- For each edge (keyword pair), compute
  - Chi-square
    - If confidence is low, delete the edge
  - Correlation Coefficient
    - If less than threshold, delete the edge
- Only strong associations remain after pruning



Segmenting the Keyword Graph

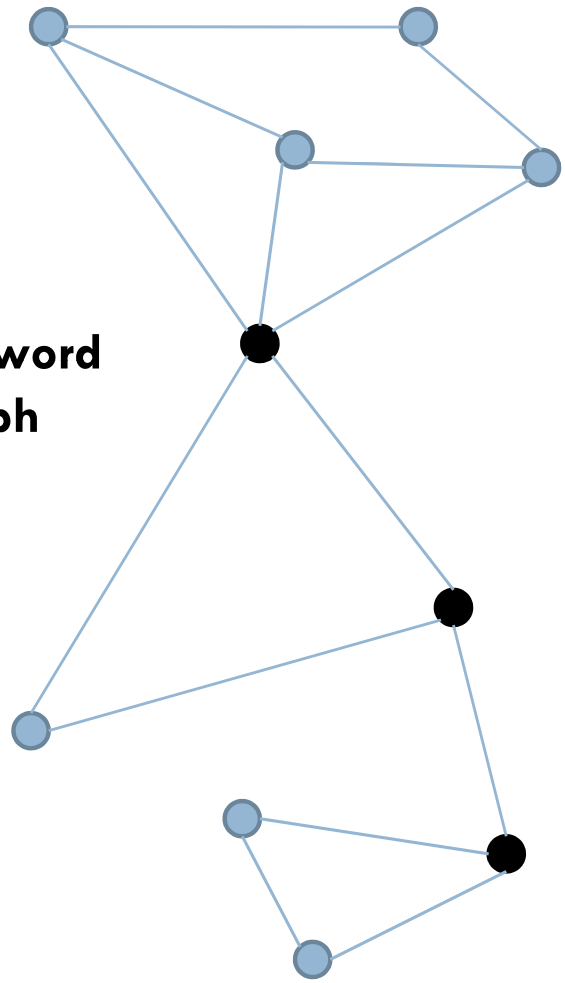
UNIVERSITY OF TORONTO

- Graph clustering algorithms [KK'98, FRT'05]
  - We don't know the number of clusters
  - High computational complexity
  - Graph may not fit in main memory
- Correlation clustering [BBC'04] - expensive
- Bi-connected components
  - An articulation point in a graph is a vertex such that its removal makes the graph disconnected. A graph with at least two edges is bi-connected if it contains no articulation points.

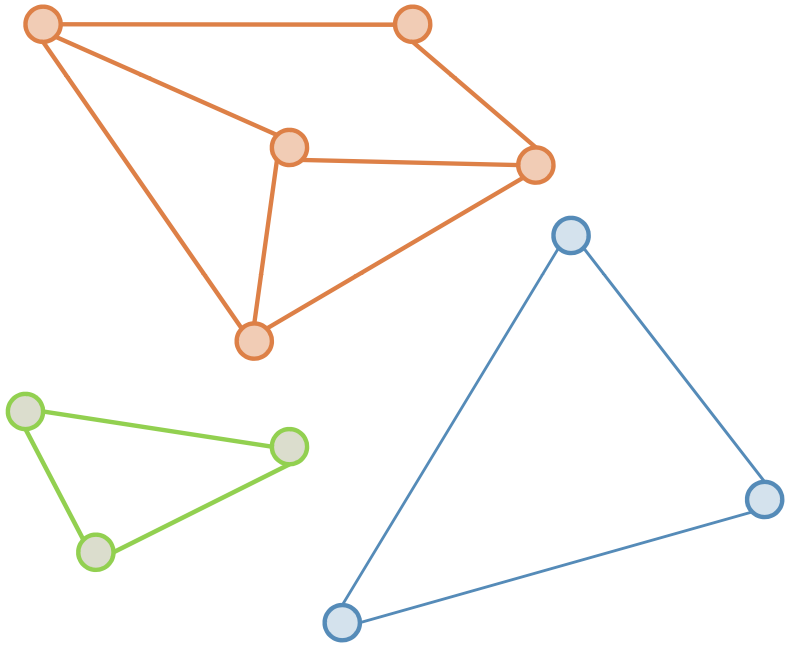
Bi-connected Components

UNIVERSITY OF TORONTO

- Segment the graph
  - Find maximal bi-connected components



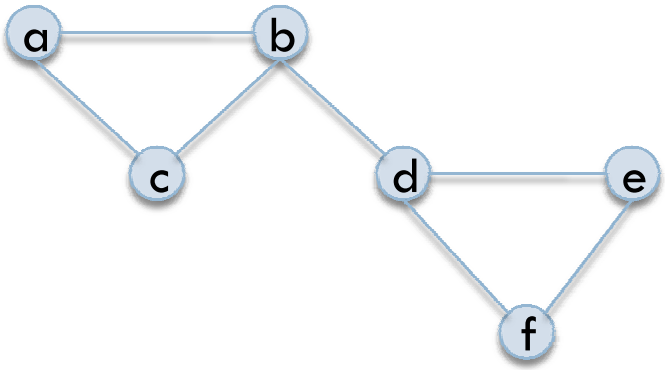
keyword graph

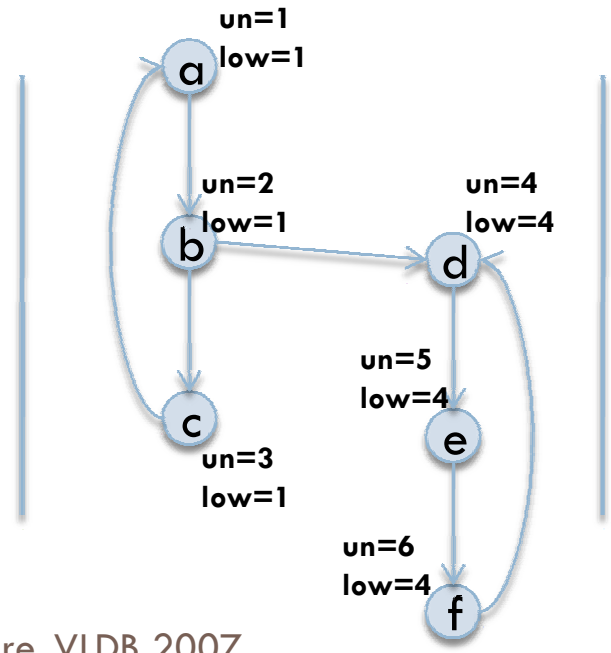


keyword clusters

Finding Bi-connected Components

- Efficient algorithm exists – single pass
  - Realizable in secondary storage [CGGTV'05]
  - Perform a DFS on the graph
    - Maintain two numbers, *un* and *low,* with each node





Bi-connected Components:

1. (f,d) (e,f) (d,e)
2. (c,a) (b,c) (a,b)

Cluster Graph

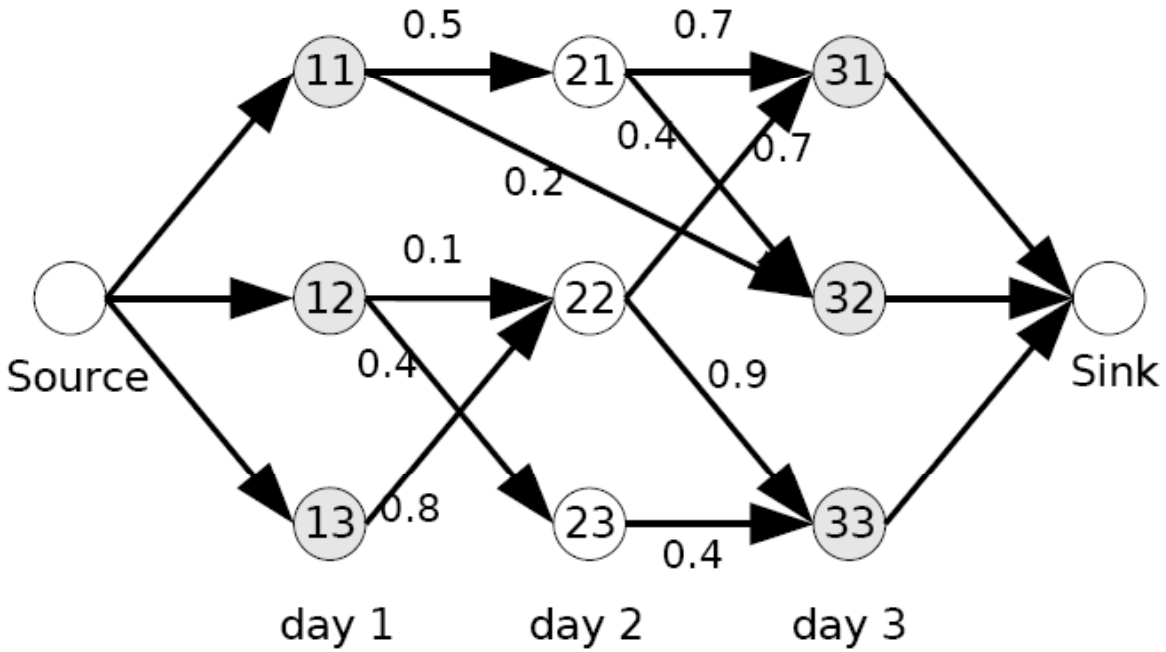
UNIVERSITY OF TORONTO

- We have a set of clusters for each time step (day)
  - Each cluster is a set of keywords
- Similarity between two clusters can be assessed
  - Intersection, i.e., number of common keywords
  - Jaccard coefficient
- Aim is to find clusters that persist over time
- A graph of clusters over time can be constructed
  - Undirected graph with edge weight equal to similarity between the keyword clusters

Example Cluster Graph

UNIVERSITY OF TORONTO

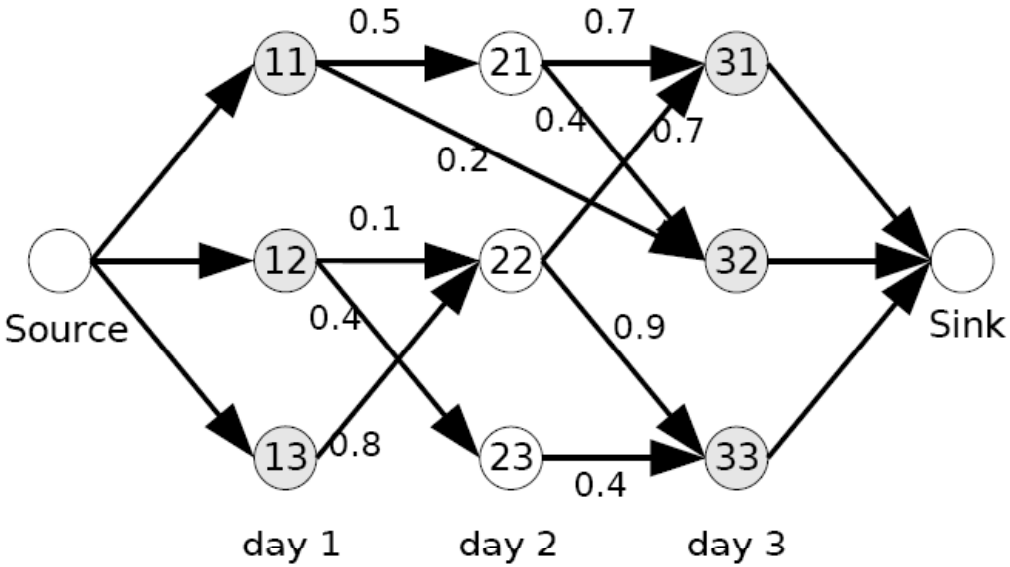
- Graph over clusters from three time steps
  - Max temporal gap size, g=1
  - Three keyword clusters on each time step
    - Each node is a keyword cluster
    - Add a dummy source and sink, and make edges directed
    - Edge weights represent similarity between clusters



Formal Problem Definitions

UNIVERSITY OF TORONTO

- Weight of path = sum of participating edge weights
- Definition: **kl-Stable clusters**
  - Find top-k paths of length l with highest weight
- Definition: **normalized stable clusters**
  - Find top-k paths of minimum length l_{min} of highest weight normalized by their lengths



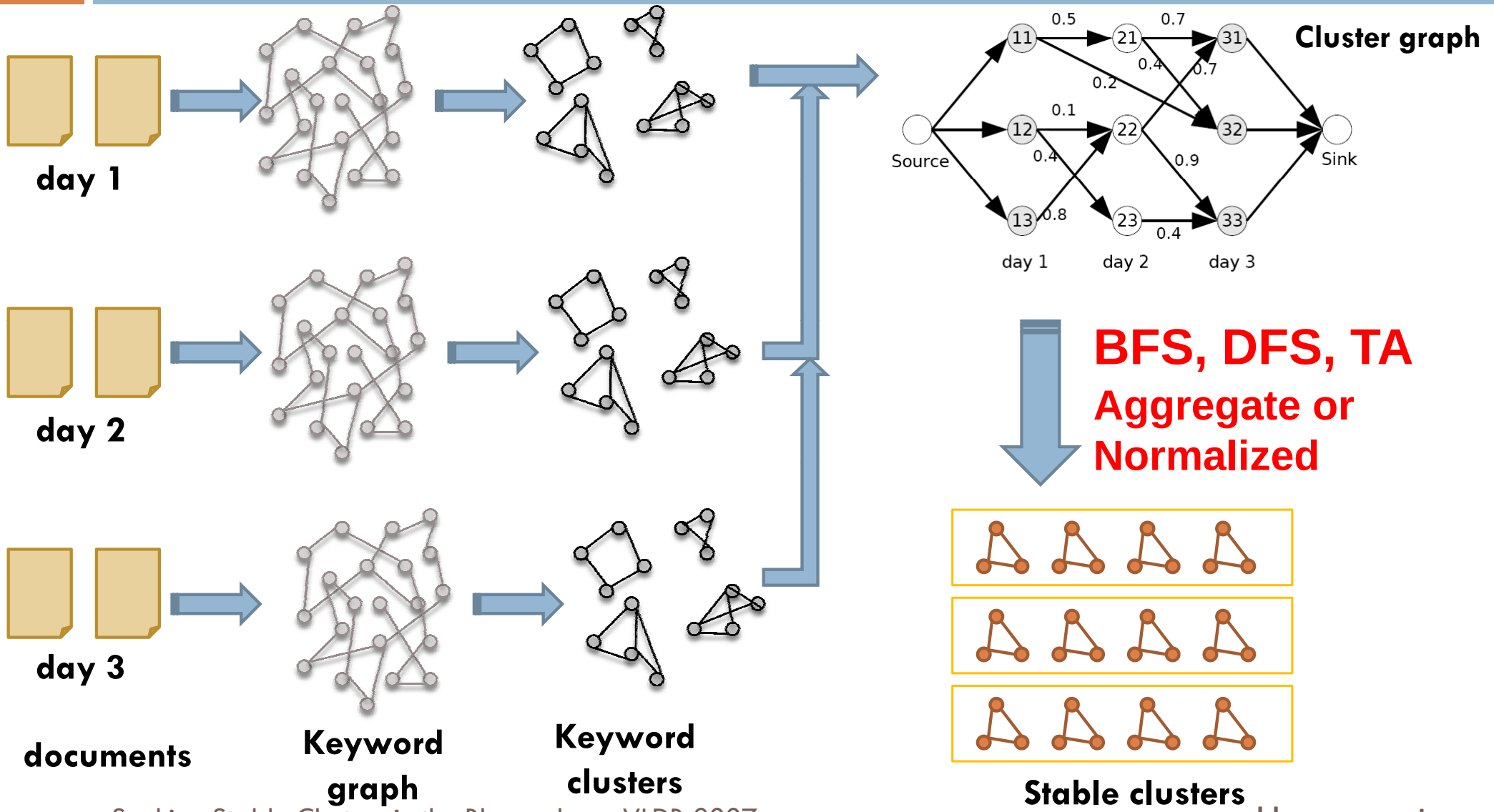
Algorithms for kl-Stable Clusters

UNIVERSITY OF TORONTO

- Breadth First Search
  - Fastest, but requires significant amounts of memory
- Depth First Search
  - Slower, but has low memory requirements
- Adaptation of the Threshold Algorithm [FLN'01]
  - Exponential number of I/Os, very slow

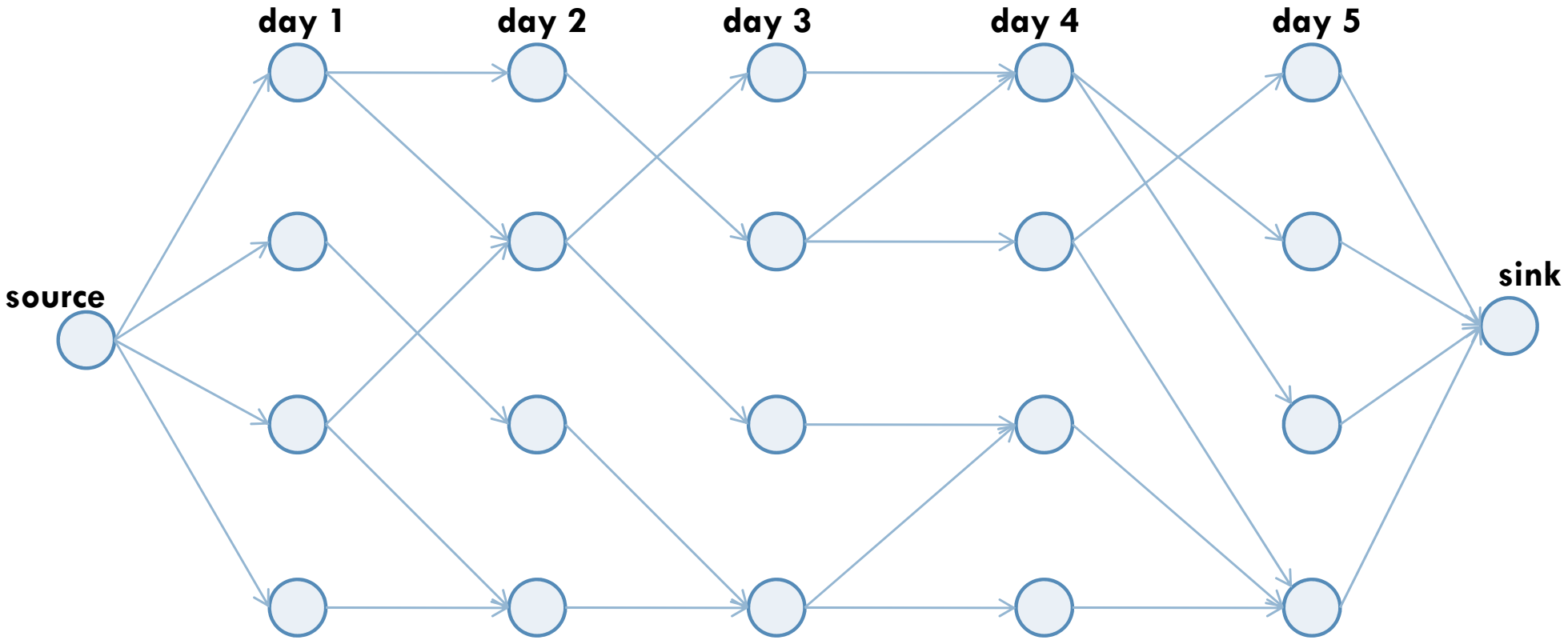
Pipeline

UNIVERSITY OF TORONTO



Breadth First Search

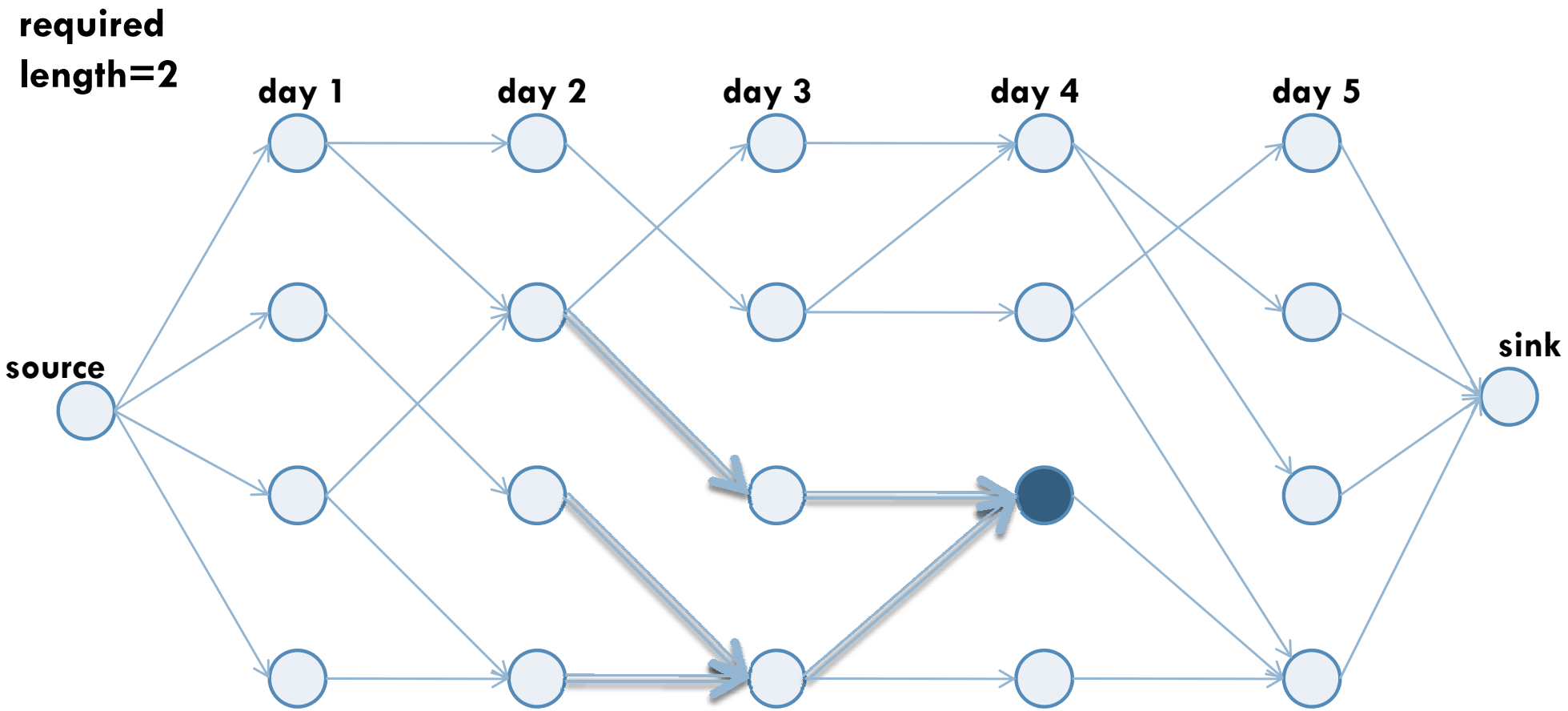
UNIVERSITY OF TORONTO



Cluster graph with max temporal gap, g=0

BFS Example

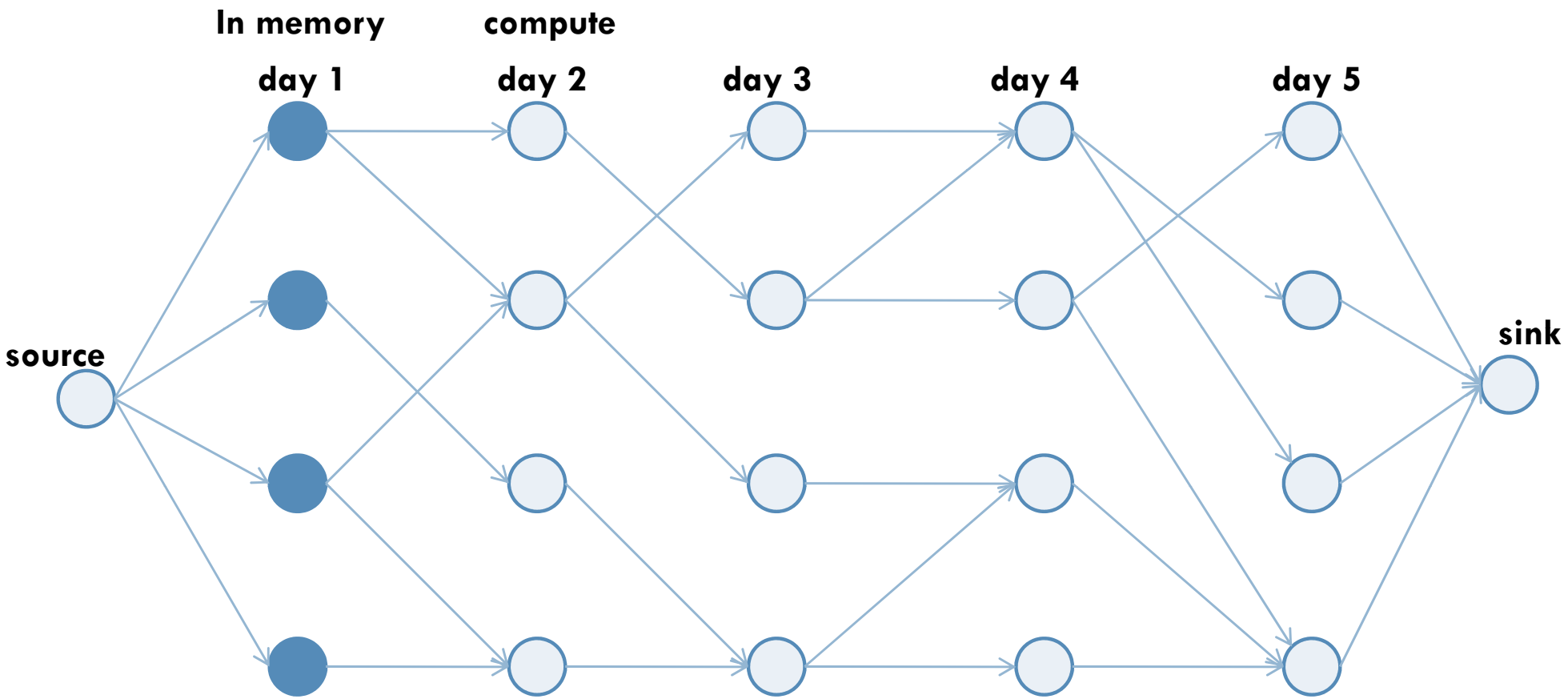
UNIVERSITY OF TORONTO



Cluster graph with max temporal gap, g=0

BFS Example

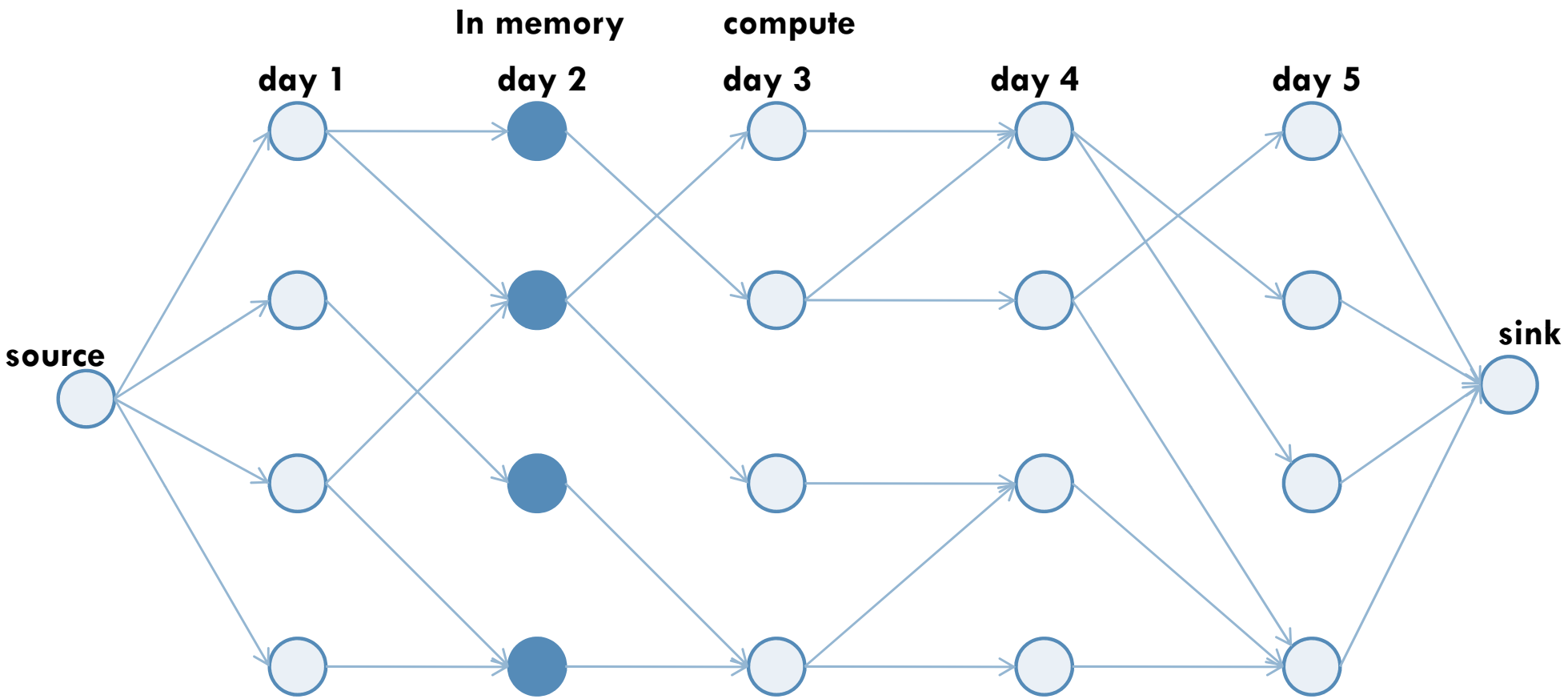
UNIVERSITY OF TORONTO



Cluster graph with max temporal gap, g=0

BFS Example

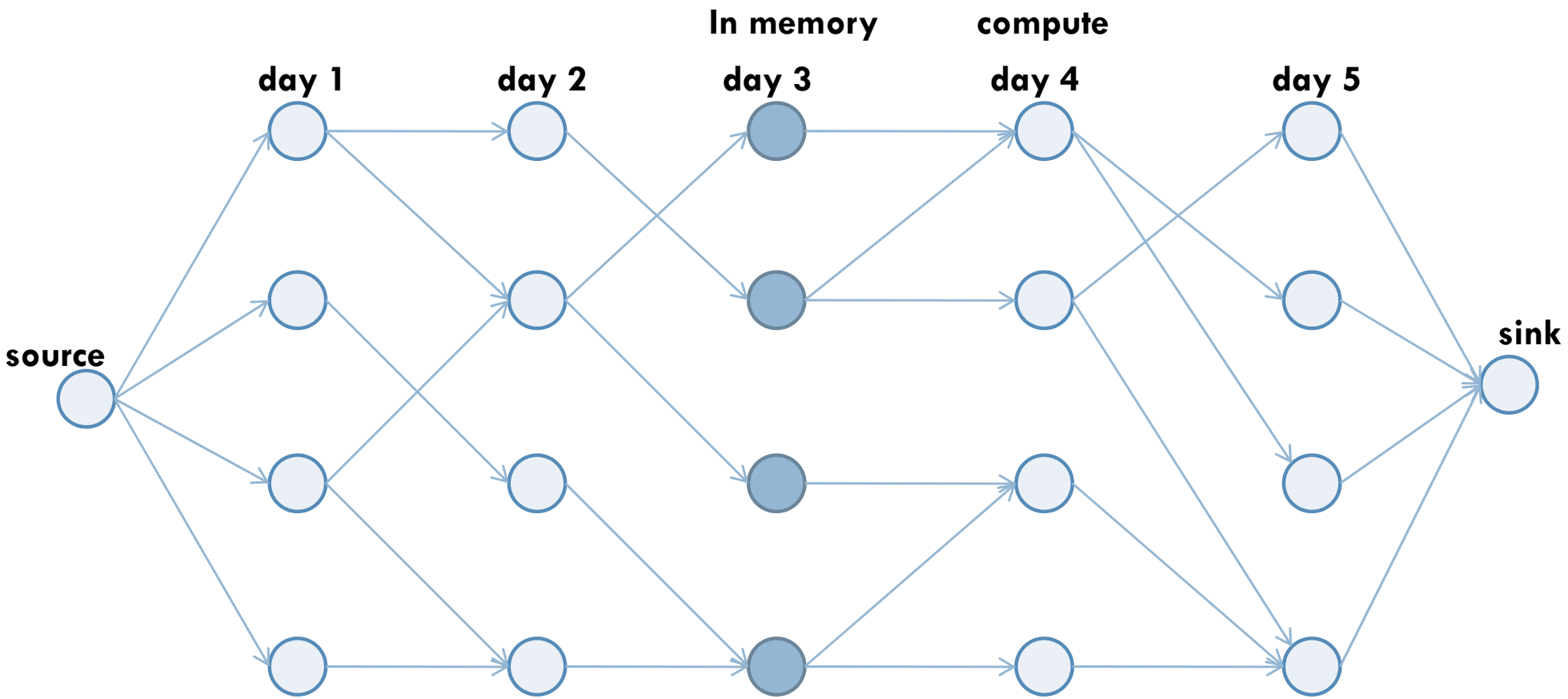
UNIVERSITY OF TORONTO



Cluster graph with max temporal gap, g=0

BFS Example

UNIVERSITY OF TORONTO



Cluster graph with max temporal gap, g=0

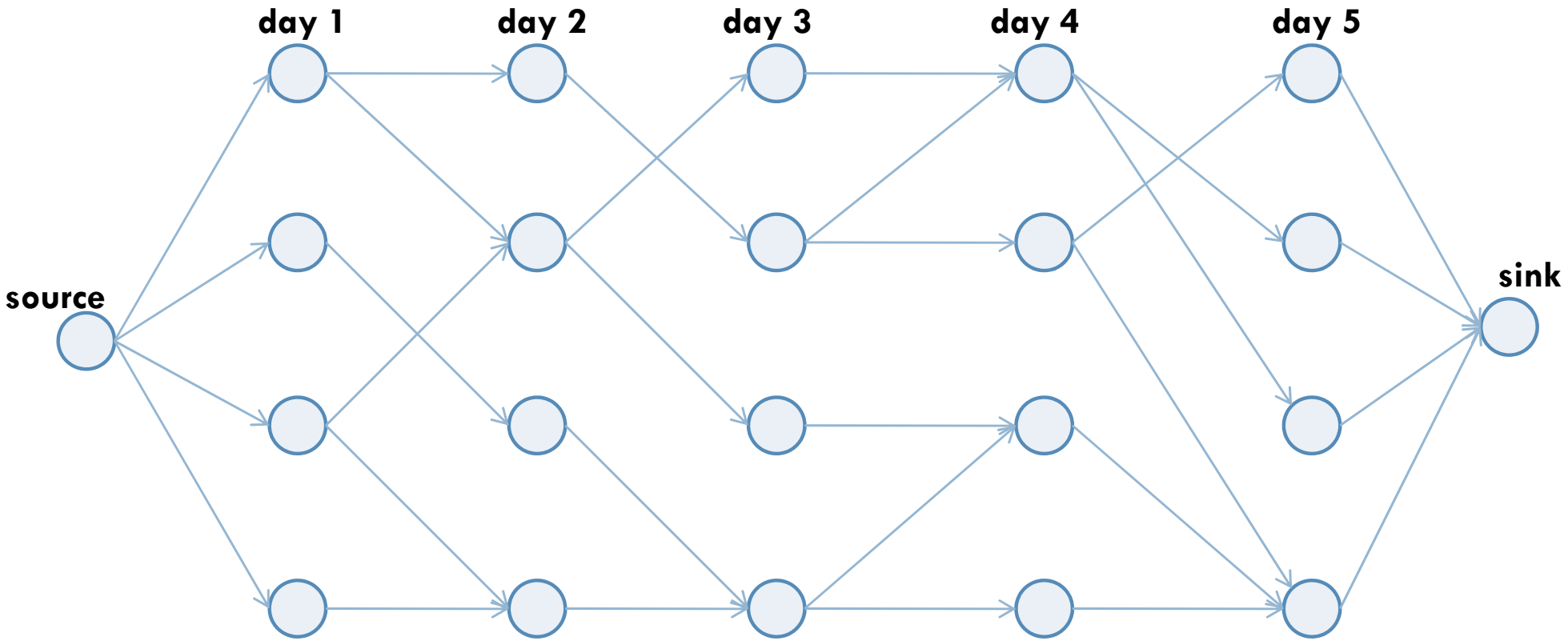
BFS Analysis

UNIVERSITY OF TORONTO

- Algorithm requires a single pass over all G_i
  - I/O linear in number of clusters (sequential I/O only)
- Needs enough memory to keep all clusters from past g+1 time steps in memory
- If enough memory is not available, multiple pass required
  - Similar to block nested join
- Amenable to streaming computation
  - Can easily update as new data arrives

Depth First Search

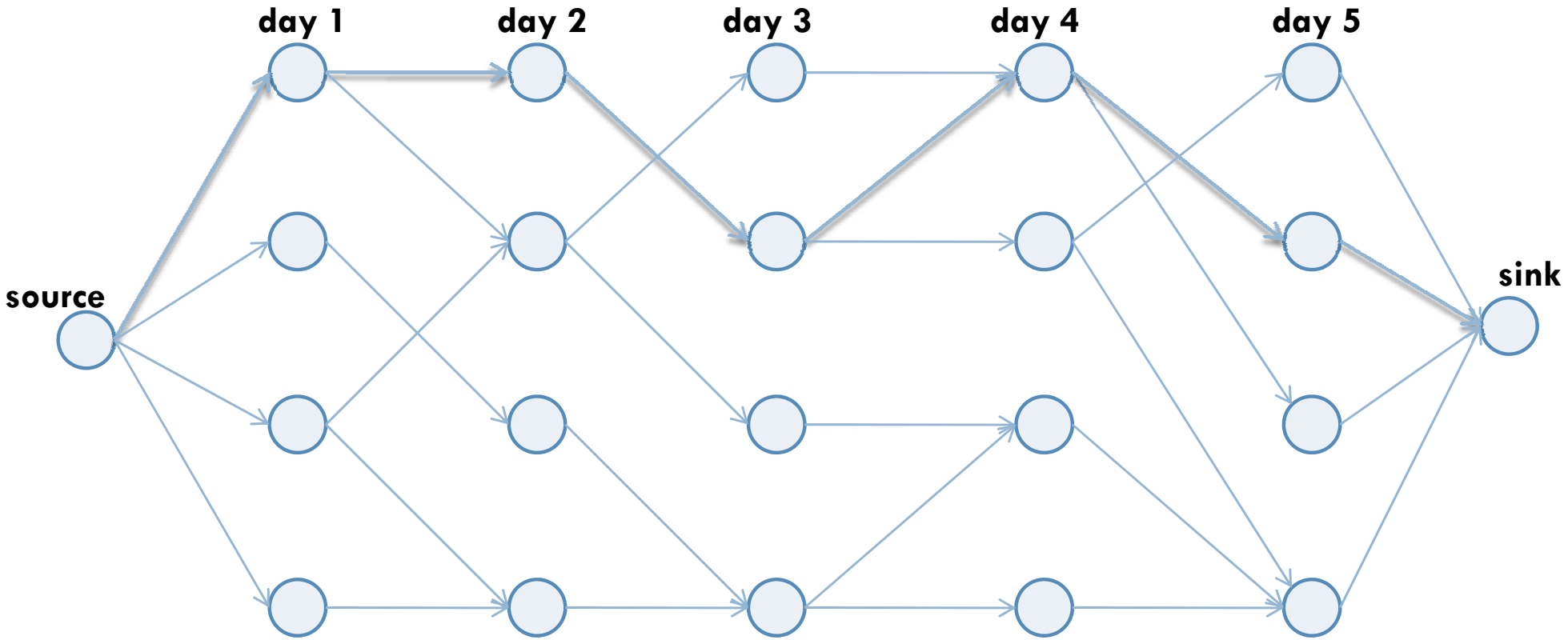
UNIVERSITY OF TORONTO



Cluster graph with max temporal gap, g=0

DFS Example

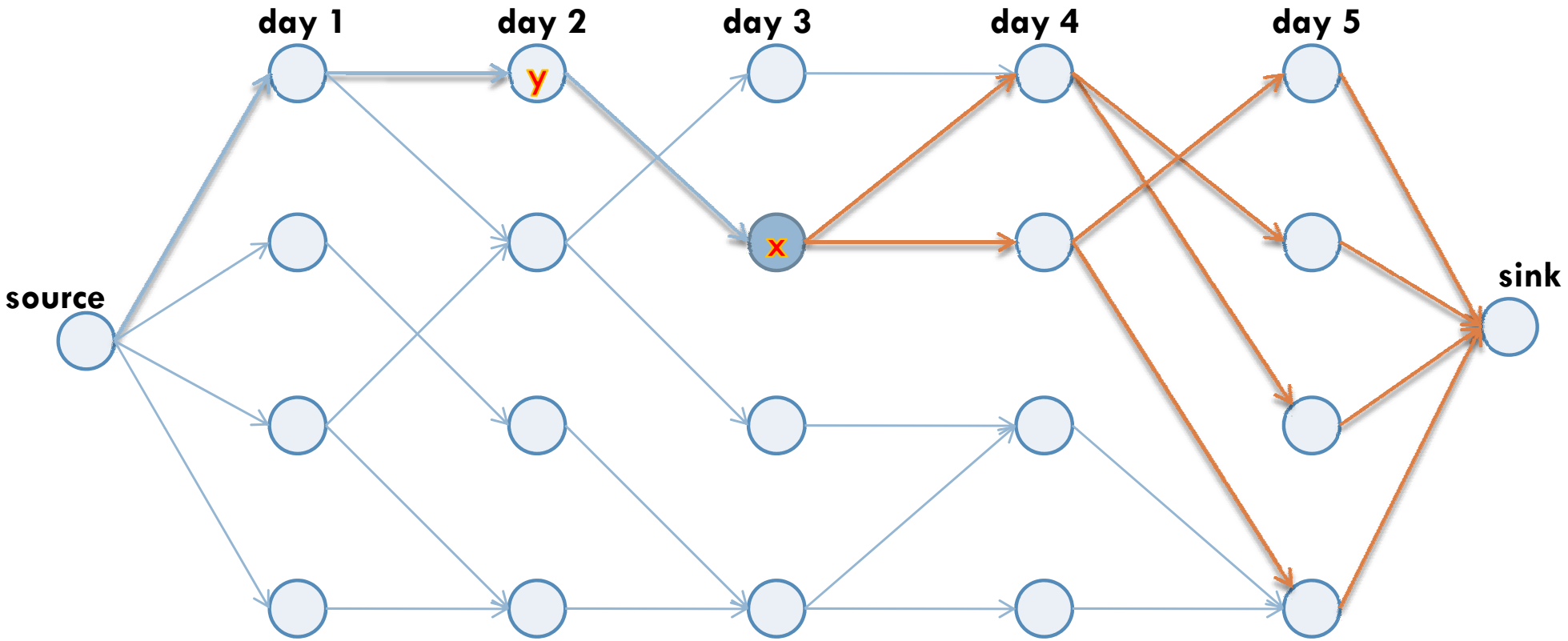
UNIVERSITY OF TORONTO



Cluster graph with max temporal gap, g=0

DFS Example

UNIVERSITY OF TORONTO



Cluster graph with max temporal gap, g=0

DFS Analysis

UNIVERSITY OF TORONTO

- The number of I/O accesses is proportional the number of edges in cluster graph
- Small memory requirement
  - Keeps the stack in the memory
  - Size of the stack bounded by total number of temporal intervals
- Can be easily updated as new data arrives

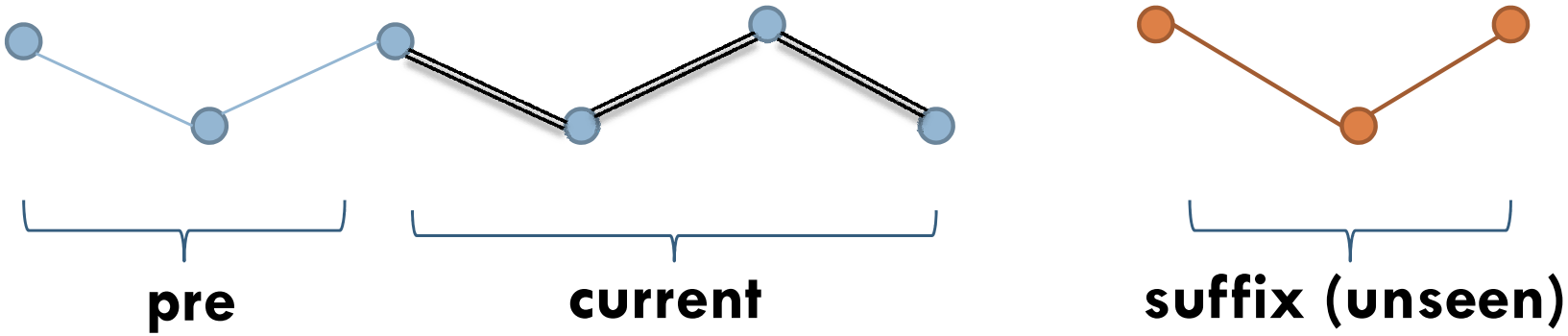
Normalized Stable Clusters

UNIVERSITY OF TORONTO

- Find top-k paths of length greater than l_{min} with highest weight normalized by their length
  - *stability*(π) = *weight*(π)/*length*(π)
- Both the BFS or DFS based techniques can be used
- Since there is no specified path length
  - Need to maintain paths of all lengths for a node
  - Increases computational complexity
- *weight*(π)/*length*(π) is not monotonic
  - Makes pruning tricky

Pruning Condition

UNIVERSITY OF TORONTO



THEOREM 1. *If* $\pi_{pre}\pi_{curr}$ *is a valid path such that,*

$$stability(\pi_{pre}) \leq stability(\pi_{curr}),$$

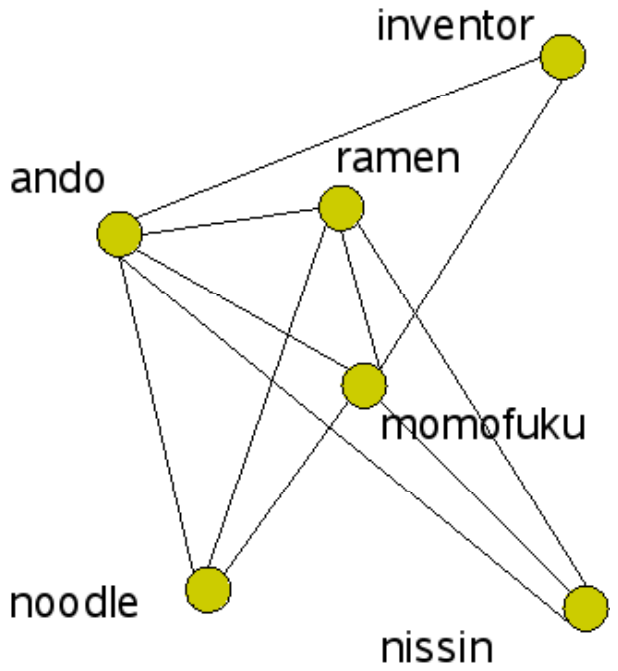
then for any possible suffix π_{suff},

$$stability(\pi_{pre}\pi_{curr}) \leq stability(\pi_{pre}\pi_{curr}\pi_{suff})$$
$$\Rightarrow stability(\pi_{pre}\pi_{curr}\pi_{suff}) \leq stability(\pi_{curr}\pi_{suff}).$$

Experiments

- We present results from blog postings in the week of Jan 6th
- Around 1100-1500 clusters were produced for each day
  - Threshold of 0.2 used for correlation coefficient

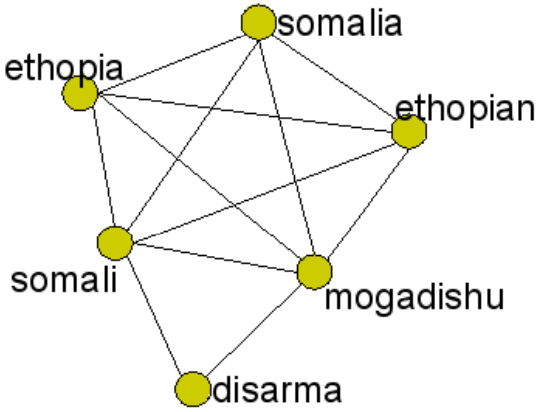
Jan 6th: Momofuku Ando, the founder-chairman of Nissin Food Products Co, who was widely known as the inventor of instant noodles, died of heart failure.



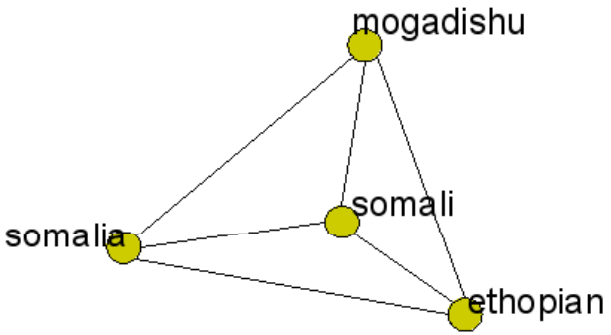
War in Somalia

The battle by Islamist militia against the Somali forces and Ethiopian troops. On Jan 9, Abdullahi Yusuf arrives in Mogadishu, and US gunships attack Al-qaeda targets.

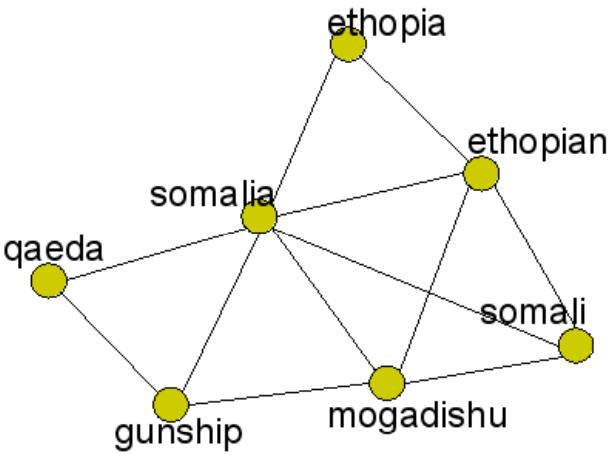
UNIVERSITY OF TORONTO



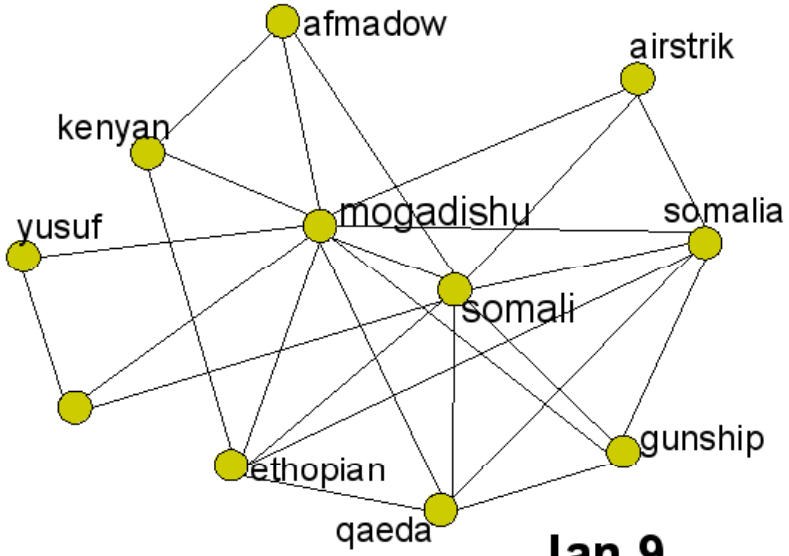
Jan 6



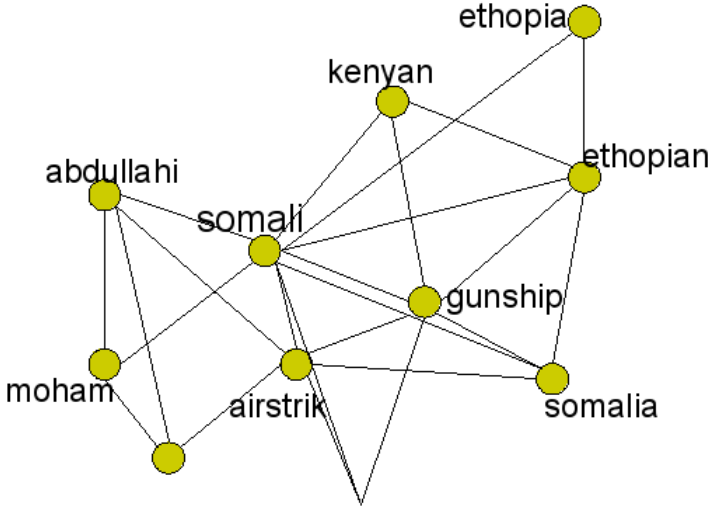
Jan 7



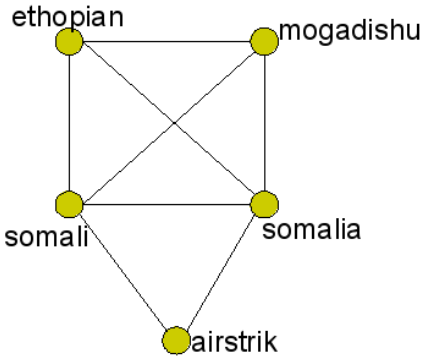
Jan 8



Jan 9



Jan 10



Jan 11

www.blogscope.net

Experiments: Performance

UNIVERSITY OF TORONTO

- Finding bi-connected components took 30 minutes when correlation coefficient threshold set to 0.2

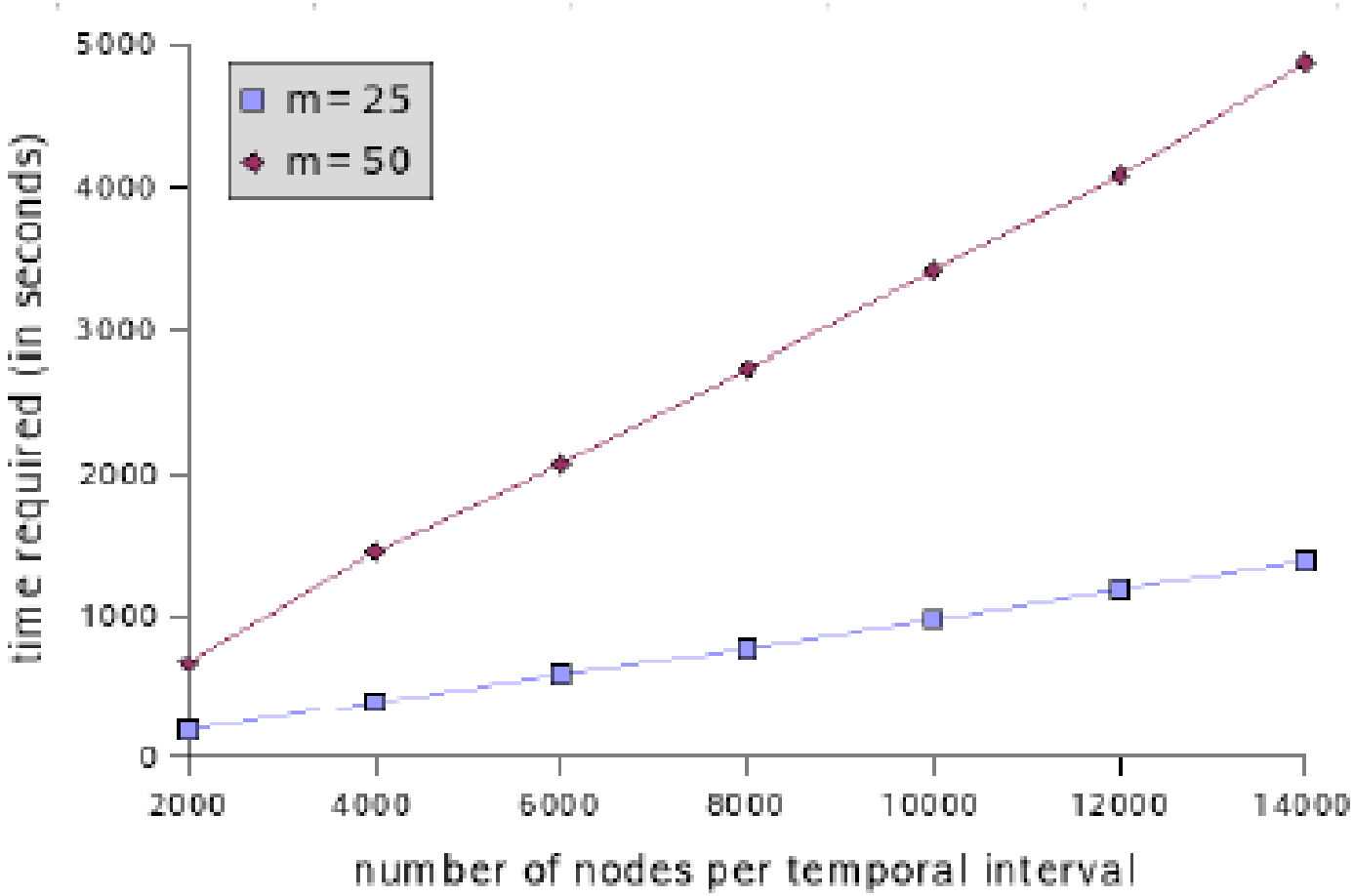
m=	3	6	9	12	15
BFS	0.65	2.09	4.49	7.95	12.49
DFS	60.3	368.8	754.8	805.94	792.05
TA	0.35	11.11	133.89	> 10 hrs	> 10 hrs

Running times on a graph with *m* time steps and 400 nodes per each time step for identifying top-5 paths.

- DFS requires less than 2 MB RAM for a graph with 2000x9 nodes, while BFS needs 35MB for the same graph.

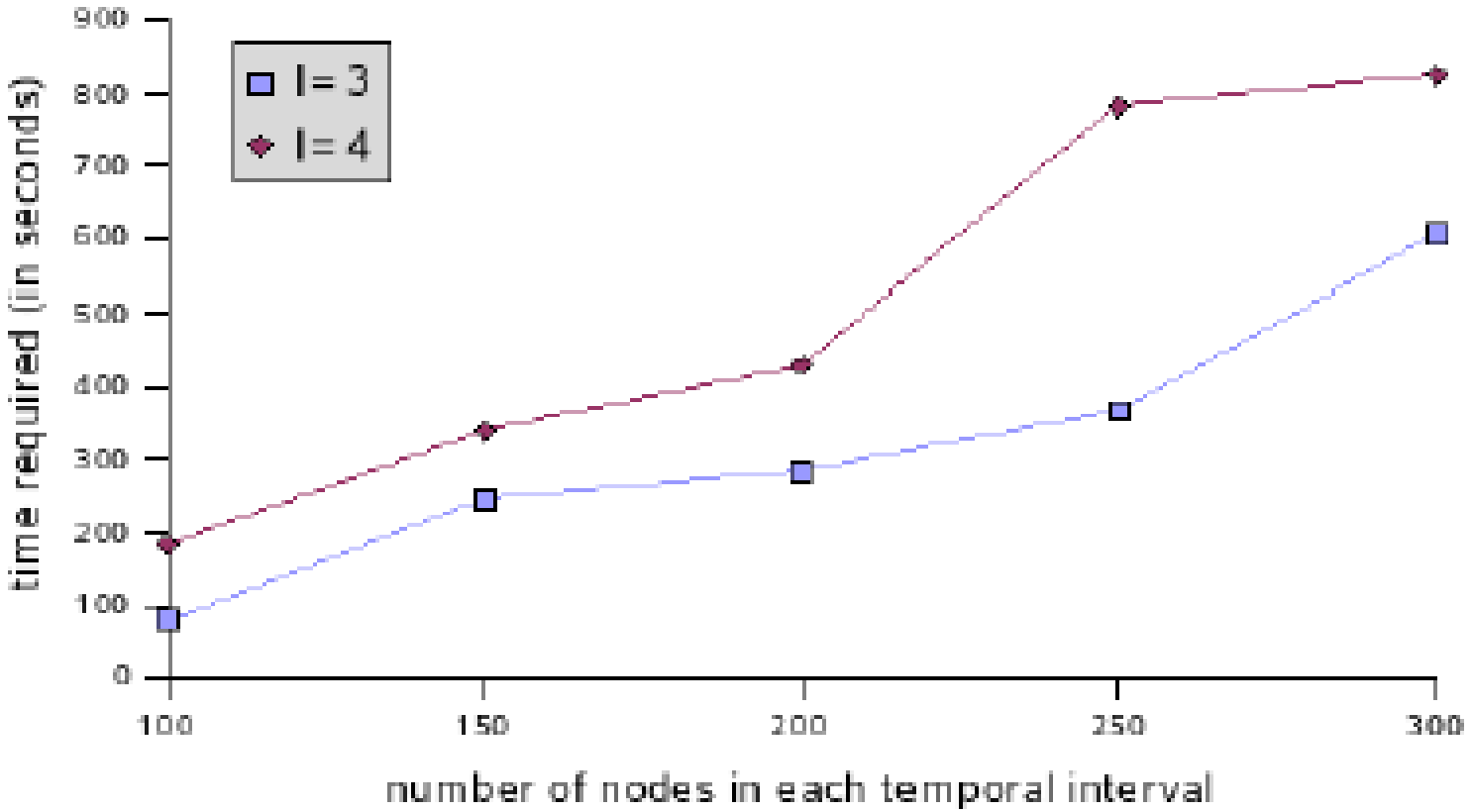
Experiments: BFS

Running time for BFS seeking top-5 paths. *m* is the number of time steps. Average out degree set to 5, and max gap size set to 1.



Experiments: DFS

UNIVERSITY OF TORONTO



Running time for DFS as we increase the number for nodes in each time step and length of the path l. Seeking top-5 path in a graph over 6 time steps

Conclusions

UNIVERSITY OF TORONTO

- Formalize the problem of discovering persistent chatter in the blogosphere
  - Applicable to other temporal text sources
- Identifying topics as keyword clusters
- Discovering stable clusters
  - Aggregate stability or normalized stability
  - 3 algorithms, based on BFS, DFS, and TA
- Experimental Evaluation

Thanks!

Visit us as **www.blogscope.net**

Nilesh Bansal, Fei Chiang, Nick Koudas, Frank Wm. Tompa

UNIVERSITY OF TORONTO

University of Waterloo