Web Graph Similarity for Anomaly Detection

Panagiotis Papadimitriou[1], Ali Dasdan[2] and Hector Garcia-Molina[1]

[1]Stanford University [2]Yahoo! Inc

Overview

Problem

- Search engines crawl the Web on a regular basis to create web graphs (∼100B vertices,∼1T edges).
- ***Anomalies:*** Factors that may result in web graphs with poor Web represenation.
- Anomalies can have a significant impact on the search results, e.g., they can affect ranking.
- Anomaly detection is difficult because of the lack of a "perfect" web graph to compare with and the large size of web graphs.

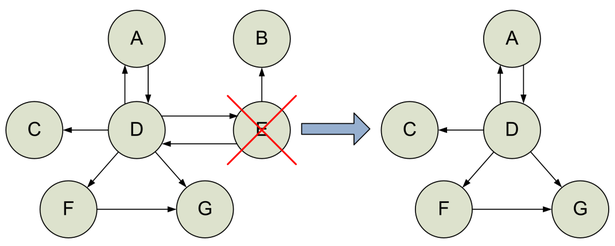
Our Approach

- We suggest detecting anomalies by measuring the amount and the significance of changes in consecutive web graphs.
- We suggest 5 similarity metrics to quantify the differences between two web graphs.
- We present an extensive experimental evaluation of our approach and the different similarity metrics.

Potential Anomalies

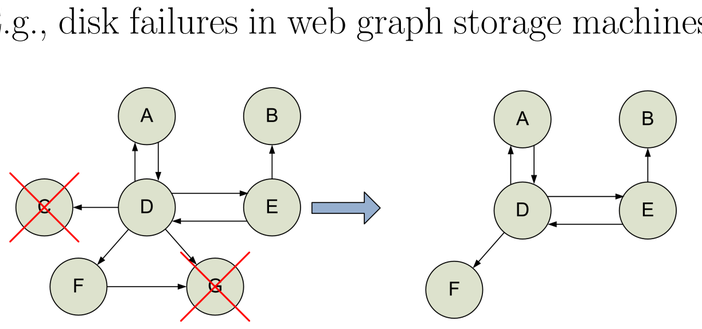
Missing Connected Subgraph

E.g., a web host is unavailable at crawl time.



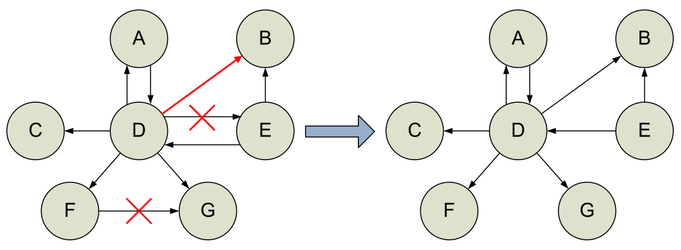
Missing Random Vertices

E.g., disk failures in web graph storage machines.



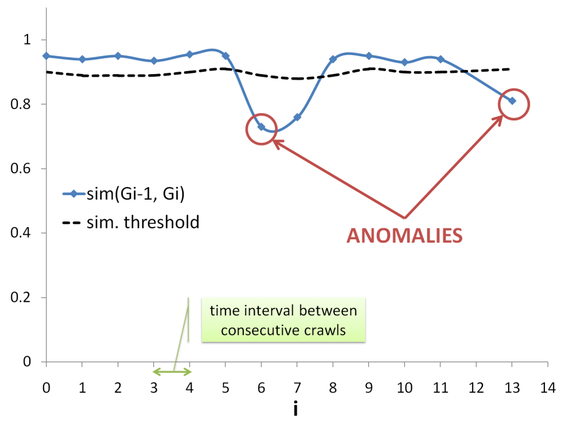
Random Topological Changes

E.g., bugs in web graph data management software result in wrong edge fetching.



Anomaly Detection

- Regular crawls yield a sequence of consecutive web graphs $G_1, \ldots, G_n$.
- Similarity scores between any graphs G_{i-1} and G_i viewed on time axis create a time series.
- Use time series to define a similarity ***threshold*** that indicates minimum similarity between consecutive anomaly-free web graphs.



Similarity Metrics

Vertex/Edge Overlap

Jaccard Index between vertex or/and edge sets.

Vector Similarity

Distance of adjacency matrices principal eigenvectors.

Vertex Ranking

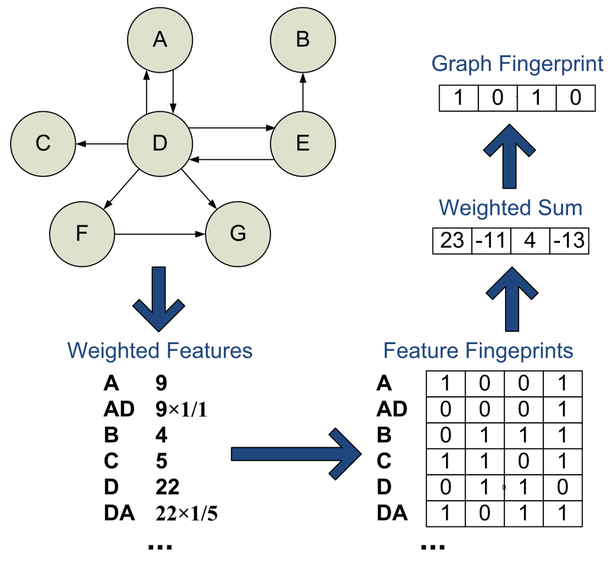
Rank correlation between sorted (by Pagerank) vertex lists.

Sequence Similarity

- Convert graphs to vertex sequences.
- Calculate ratio of common subsequences [1].

Signature Similarity

- Get graphs fingerprints using LSH [2, 3].
- Calculate Hamming distance-based similarity.



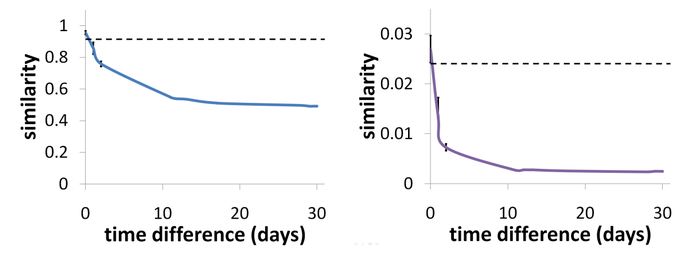
Web Graph LSH function.

Experimental Results

We used real web graphs generated by Yahoo! over a month long period in 2007.

Definition of Similarity Threshold

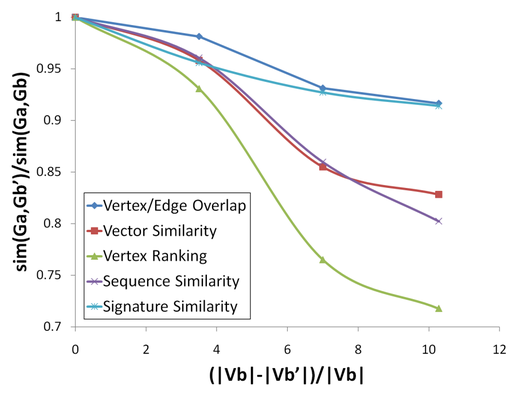
We studied how web graph similarity varies over time for the different metrics to define a similarity threshold for each metric.



Vertex/Edge Overlap. Sequence Similarity.

Similarity Metrics Evaluation

We selected consecutive graphs Ga, Gb and simulated anomalies to Gb to create corrupted versions Gb'. We used all different similarity metrics and tried to identify the anomalies using the similarity score between Ga, Gb' and the corresponding similarity threshold of each algorithm.



Missing Connected Subgraph Simulation

Signature Similarity outperforms other metrics.

	Missing Subgraph	Missing Vertices	Topological Changes
Vertex/Edge Overlap	✓✓	✓	×
Vector Similarity	✓✓	✓	✓
Vertex Ranking	✓✓	×	×
Sequence Similarity	✓	×	✓✓
Signature Similarity	✓✓	✓✓	✓✓

References

[1] A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic clustering of the web. In *Proc. of Int. World Wide Web Conf. (WWW)*, pages 393–404, Apr 1997.

[2] M. Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. of Symp. on Theory of Comput. (STOC)*, pages 380–388. ACM, May 2002.

[3] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina. Web graph similarity for anomaly detection. Technical Report 2008-1, Stanford University, 2008. URL: http://dbpubs.stanford.edu/pub/2008-1.