The Filter-Placement Problem and its Application to Content De-Duplication

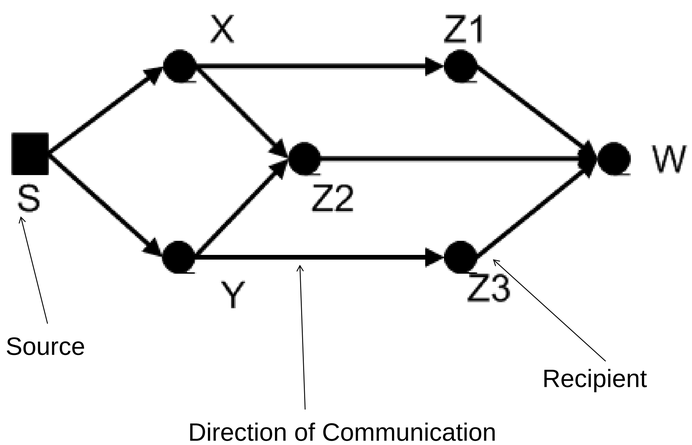
Dora Erdos, Vatche Ishakian, Andrei Lapets, Azer Bestavros, and Evimaria Terzi

Abstract

In many networks, (e.g., RSS feeds, blogs, sensor networks, ad-hoc networks) nodes blindly relay information they receive to neighbors. This uncoordinated data dissemination often results in significant, yet unnecessary, communication and processing overheads and reduce the utility of the network. To alleviate the negative impacts of information multiplicity, we propose that a subset of nodes (selected at key positions in the network) carry out additional information de-duplication functionality. We refer to such nodes as **filters**. We formally define the **Filter Placement** problem as a combinatorial optimization problem, and study its computational complexity for different types of graphs. We also present polynomial-time approximation algorithms for the problem. Our experimental results indicate that less than a handful of filters are enough to alleviate more than 95% of the redundant information.

Propagation Model

- Communication networks can be represented by directed graphs. The nodes correspond to actors in the network, and directed edges indicate the direction of information flow.
- A source in the network generates items.
- When an actor receives an item, it will make copies and propagate a copy of the item to every child of his.
- Every item may be viewed as if it travels a path from the source to a given node. Hence a node may receive several copies of the same item, one copy along every path leading from the source to that node.



Goal: Content de-duplication achieved by judicious placement of filters

Example Applications

. News Media Networks
- Reduce number of duplicate syndicated news items

. Networks of RSS-feeds
- Remove redundant feeds

. Sensor Networks
- Remove duplicate query answers
- Reduce energy needed to exchange duplicate measurements

. Content Networks
- Remove duplicate updates
- Effective eavesdropping

Optimum Filter Placement

FP problem: Given a directed graph G(V,E) and an integer k, find a set of nodes A in V of size k, so as to **maximize** the gain function

$$F(A) = \Phi(0,V) - \Phi(A,V)$$

Number of items received in the original network

Number of items received with A being aggregators

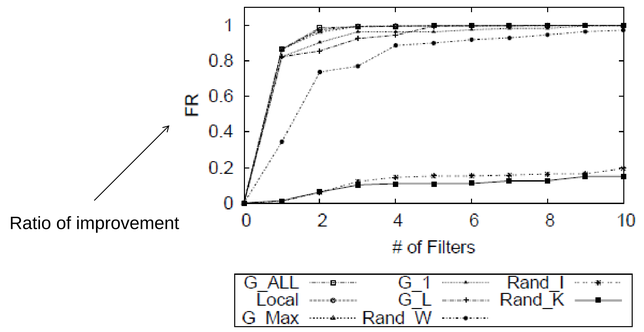
Results

- FP is NP hard on arbitrary graphs.
- FP can be solved in polynomial time with dynamic programming on trees.
- FP is even NP hard on DAGs. However a greedy algorithm can achieve an (1-1/e)-approximation.
- Various heuristics for FP turn out to be faster and in practice as effective as our greedy algorithm.
- An optimal solution can be achieved by placing filters on all nodes.

Experiments

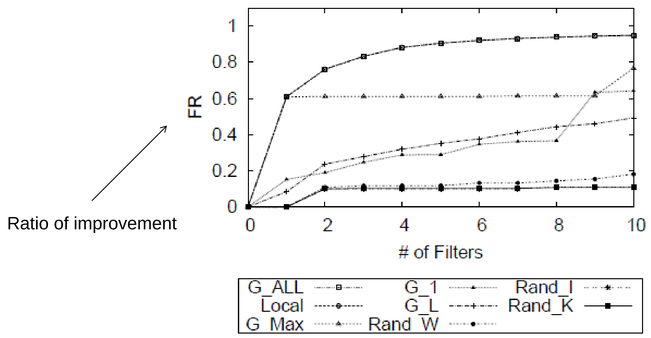
News Dataset:

Graph of the online media network from large news outlets to micro blogs. In particular, it follows the spread of the phrase *"Lipstick on a pig"* during the 2008 presidential campaign.



Physics Reviews dataset:

Graph representing the citation network of physics papers for over 100 years. Data portrays the propagation of an original concept or idea, represented by the paper at the source, through the physics community.



Bottom Line:

In typical information networks, as few as 3 well-placed filters are enough to remove as much as 90% of duplicate information!

Reference

[1] Azer Bestavros, Dora Erdos, Vatche Ishakian, Andrei Lapets, Evimaria Terzi, The filter-placement problem and its application to content de-duplication. . Tech Report 2011-005, BU, CS Dept Feb, 2011