SAND2018-2042C

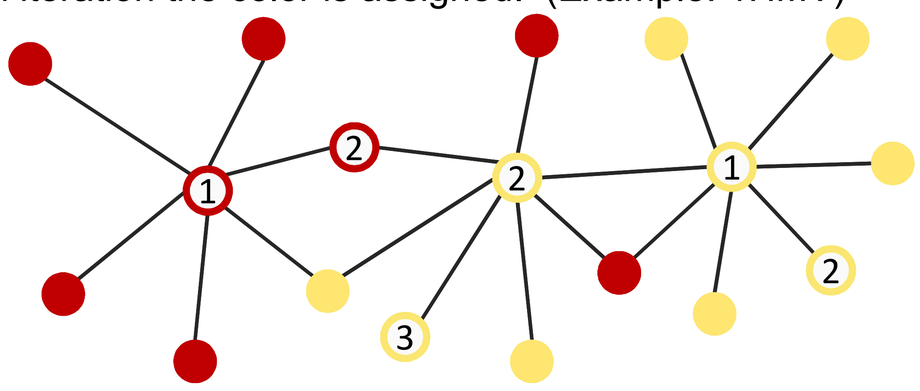
Counter-Adversarial Node Labeling

Sandia National Laboratories
W. Phillip Kegelmeyer, Jeremy D. Wendt, Ali Pinar, Cliff Anderson-Bergman
{wpk, jdwendt, apinar, ciande}@sandia.gov
Sandia National Laboratories, California 94551

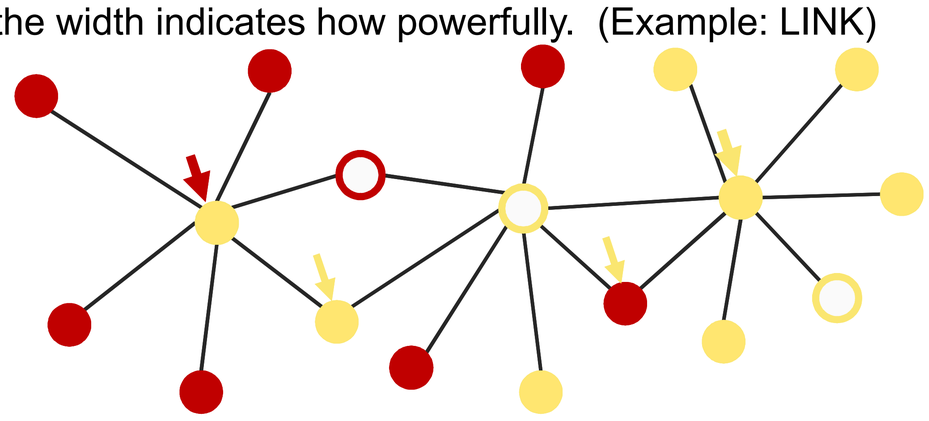
Problem

Social media enables easy, long-term communication with friends. Most of these systems allow users to specify which personal information to share or hide. Nonetheless, hidden information can often be inferred using node labeling techniques.

Voting Algorithms assign a node's neighbors' majority label (red or yellow) to each unlabeled node (white center) – iterating until node labels converge (outer line color). Numbers indicate in which iteration the color is assigned. (Example: 1HMV)



Witness Algorithms use supervised machine learning to exploit "witness nodes" that predict their neighbors color based on colors of labeled neighbors. Below, the arrows' color indicates what color the node witnesses (most of its neighbors this color), and the width indicates how powerfully. (Example: LINK)

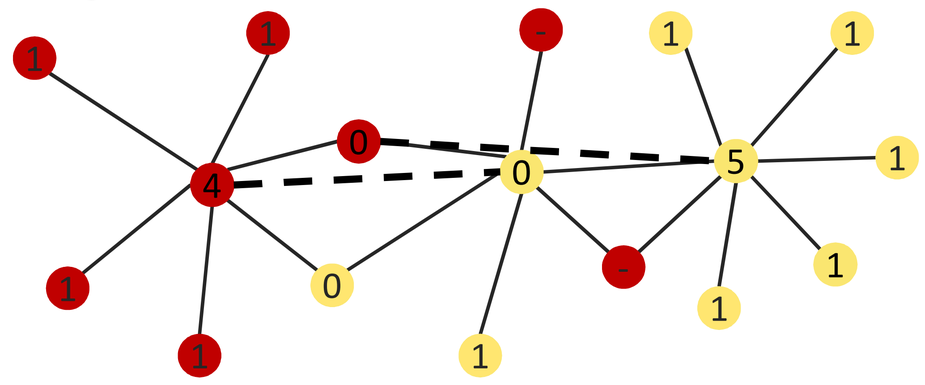


Attacks

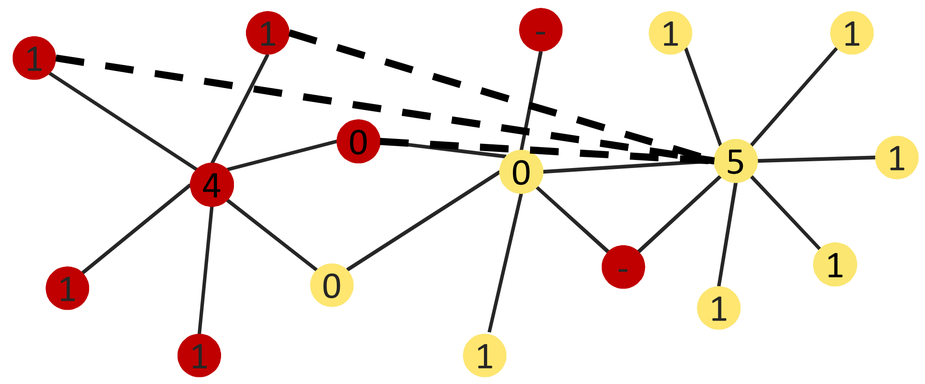
We allow the attacker to add edges between any two nodes in the graph. His goal is to reduce overall accuracy as much as possible.

Inspired by voting algorithms, we define each node's *conversion cost* (***cc***) as the difference between its same-color neighbors and its different-color neighbors. (All nodes with negative ***cc*** are already mislabeled by voting algorithms.) We invented two heuristics to decrease accuracy based on ***cc***.

The Weak-willed Attack adds edges from low-***cc*** nodes of each label to high-***cc*** nodes of the other label.



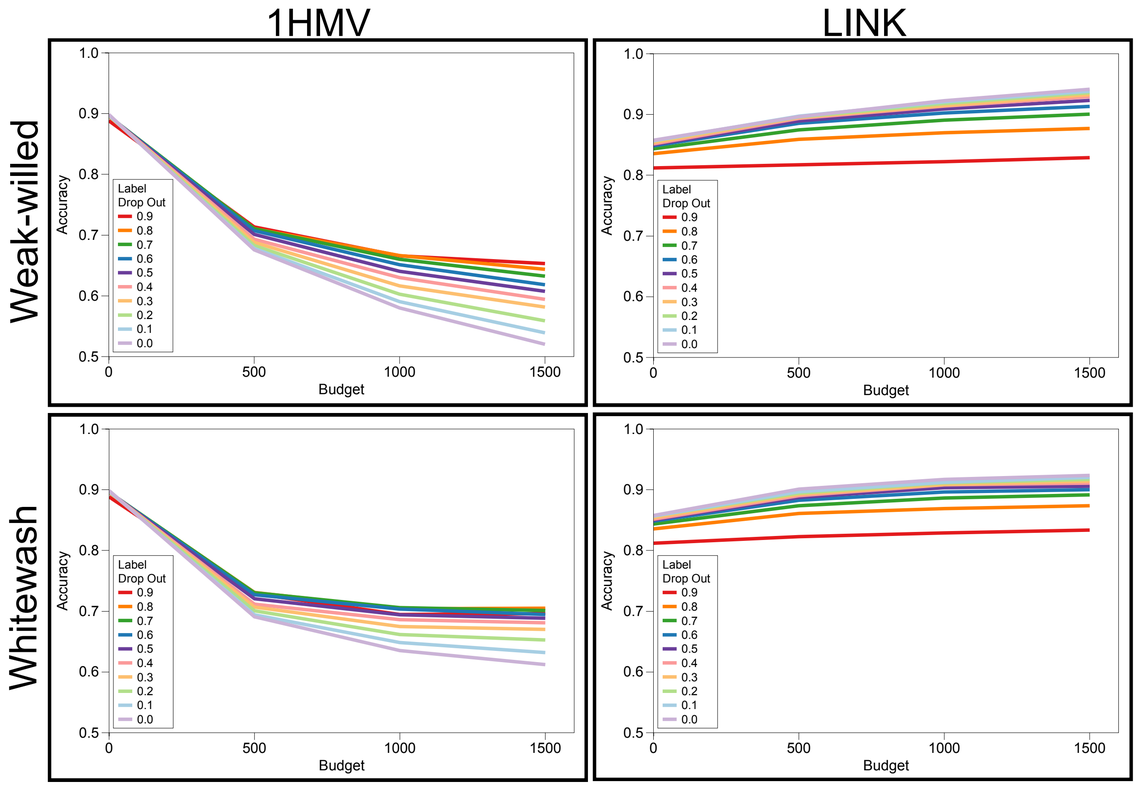
The Whitewash Attack adds edges from low-***cc*** nodes of the minority label to high-***cc*** nodes of the other label.



Results

We used the 2004 Political Blogs Dataset (1,490 nodes; 19,025 edges; 50.9%/49.1% label split). We made 20 instances of the weak-willed and whitewash attacks (different random seeds). On the right we see the averaged results of 10-fold cross-validation of these runs.

"Label Drop Out" indicates what percentage of the 9-folds of labeled nodes were dropped before inferring the 1-fold held-out nodes' labels.



Our attack heuristics degrade 1HMV accuracy; LINK accuracy increases – undesired and surprising. This may be due to the attacks adding edges between many low ***cc*** nodes and specific high ***cc*** nodes of the other label class – creating a pattern for the LINK's ML to exploit. If a node has edges to these high ***cc*** nodes, but not many other nodes in that label class, LINK can learn these nodes are of their original color.

We are developing LINK-attacking heuristics and hope they will degrade both methods' accuracy.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

Social media image thanks: https://commons.wikimedia.org/wiki/File:Social_media_icon.png