Network Sampling Designs for Relational Classification

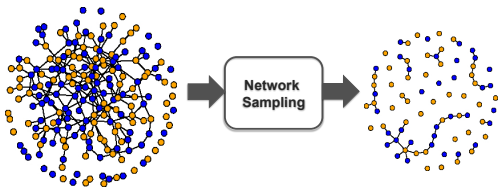
Nesreen K. Ahmed, Jennifer Neville, and Ramana Kompella
CS Department, Purdue University

Introduction

- Online social networks (OSNs) have witnessed a huge popularity recently (e.g. Facebook, MySpace, and Twitter)
- Studying complex networks is a challenging task, due to their:
  - Heterogeneous and dependent structure
  - Large size
  - Evolution over the time
- **Network Sampling** is a standard approach to select a subset of nodes/edges from the full network
  - E.g. Node Sampling, Edge Sampling, Forest Fire Sampling
- Research focused on *how to collect samples that closely match topological properties of the full network.* (e.g. Degree distribution, Clustering)

Topological properties are never entirely preserved via sampling

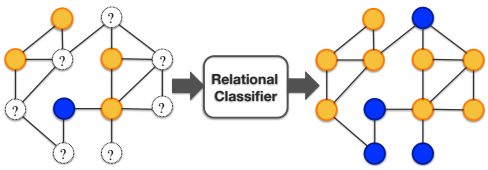
As a result, Network Sampling impacts the performance of applications overlaid on the network



Full Network **Sampled Network**

Problem Definition

- There has been a little focus on how the sampling process impacts the performance of applications overlaid on the network
- **Goal:** We study the challenging question of how the choice of sampling design impacts the evaluation of performance of relational classification algorithms
- Due to missing nodes/links, network sampling can produce samples with:
  - Imbalance in class membership
  - Bias in topological features (relational features) due to missing nodes/edges
- Relational classification attempts to identify the unknown class to which an entity belongs based on a training set of dependent entities
  - (e.g. identifying political views of connected friends in Facebook)



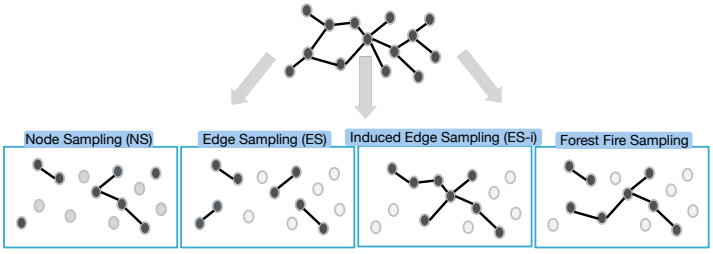
Input Network (with few known labels) **Output Network**

- Weighted-vote relational neighbor (wvRN) *(Macskassy et. al, 2007)*
  - Classifying a node using class labels of known neighbors

Classes of Sampling Designs

- Samples are constructed by selecting a subset of nodes/edges randomly
- A sample is considered representative of the original graph if it preserves key characteristics of the graph. (e.g. degree, path length, clustering coefficient)
- Network Sampling can be broadly classified as:

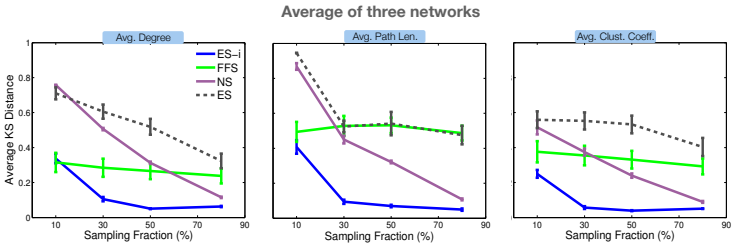
(1)Node-based (2) Edge-based (3) Topology-based



Experiments & Results

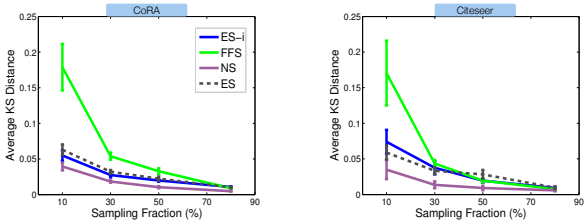
- **Data**:
  - Two citation networks (CoRA, Citeseer)— 2708, 3312 nodes respectively
  - Facebook friendship graph from Purdue network — 7315 users
- **Methodology**:
  - Compare the classification accuracy of the sampled network to the classification accuracy of the full network
  - Algorithms are evaluated on degree, path Length, and clustering coefficient
  - Samples are generated between 10-80% of original size, 10 repeated experiments for each one
  - Proportion of initially known class labels 10-80% selected randomly from the network.
  - Using 5-cross validation for evaluation

Topological Graph Properties



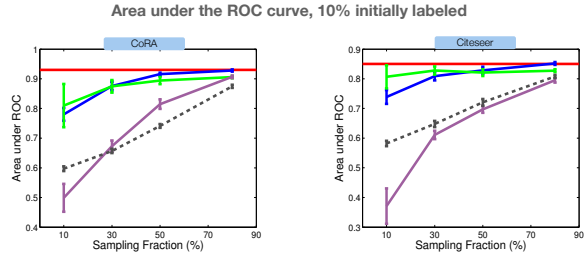
- *NS and ES produce sparse samples (don't match topological properties of the full network)*
- *ES-i and FFS outperform the other methods*

Class Distribution



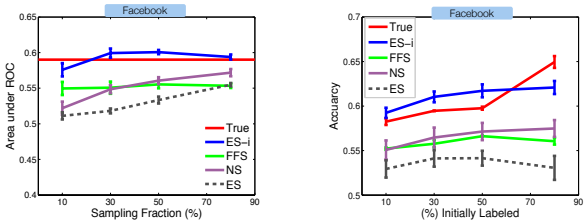
- *NS, ES, and ES-i samples closely match the original class distribution*
- *FFS samples have a large bias at 10% sampling rate (imbalance class membership)*

Classification Accuracy



- *Classification accuracy (AUC) is underestimated for sample sizes < 50%*
- *ES-i and FFS produce estimates of "AUC" close to the "True" AUC on the full network*
- *Methods that match topological properties of the full network also produce good estimates of "True" accuracy on the full network*

Accuracy for Facebook Network



- *ES-i estimates of "AUC" converge to the "True" AUC on the full network for different percentage of initially labeled nodes*