# Contents

**1 problem description** **1**

**2 methodologies** **1**
2.1 random walks . . . 1
2.2 the topic model in our method . . . 1

**3 related models** **1**

**4 stochastic variational inference** **1**

**5 experiment** **1**
5.1 evaluation metrics . . . 1
5.2 data sets . . . 2

# 1 problem description

Network community detection.

# 2 methodologies

Randoms + Bayes Nonparametric Topic Model

## 2.1 random walks

treat vertexes as words, random walks as documents

## 2.2 the topic model in our method

HDP topic model

# 3 related models

SIP2-LDA:

BCD:

$$\begin{aligned} z &\sim CRP(\alpha) && clusterassingment \\ \eta_{lm} &\sim Beta(\beta, \beta) && linkprobability \\ A_{ij} &\sim Bernoulli(\eta_{z_i z_j}) && link \end{aligned}$$

Walktrap:

# 4 stochastic variational inference

# 5 experiment

## 5.1 evaluation metrics

Given a subset $S$ of $V$, let $(S, S(E))$ be the subgraph induced by $S$. Let $n_S$ be the size of $S$, $m_S$ be the number of edges inside $S$, and $c_S$ be the number of edges with one end in $S$ and the other outside $S$.

1. internal density: $D = \frac{2m_S}{n_S(n_S-1)}$
2. cut ratio: $CR = \frac{c_S}{n_S(n-n_S)}$
3. conductance: $C = \frac{c_S}{2m_S+c_S}$
4. modularity: $Q = \sum\limits_{i=1} C(e_{ii} - a_i^2)$, where $e_{ij}$ is the fraction of edges with one end in community $i$ and the other in community $j$, $a_i = \sum_j e_{ij}$. This index falls in $[-0.5, 1)$. The larger the better. Modularity is the fraction of edges that fall within the given groups minus the expected fraction if edges were distributed at random.
5. perplexity: $\exp\{-\frac{\sum\limits_{d=1}^{M} \log w_d}{\sum\limits_{d=1}^{M} N_d}\}$, the exponential of the negative average log-likelihood or the geometric mean of $1/\log_i$. The lower the better.

## 5.2 data sets

Currently our method scales to network with million nodes and achieves highest performance compared to other generative models. It also outperforms other non-probabilistic based network community detection method such as *Walktrap*.