Combining random walks and nonparametric topic model for network community detection

Ruimin Zhu

Northwestern University

ruiminzhu2014@u.northwestern.edu

May 2, 2017

Outline

1. Review of network community detection
2. Inspirations
   - random walk + deep learning
   - SSN-LDA
3. RW-HDP
   - basic idea
   - Random walks
   - HDP topic model
   - Inference
   - Community assignment
4. Experiments
   - data sets
   - evaluation metrics
   - comparison models
   - results
   - Pros and Cons
5. Future works

Network community detection

- matrix factorization methods
- optimization methods
- generative models
- other methods

Where do the inspirations come from

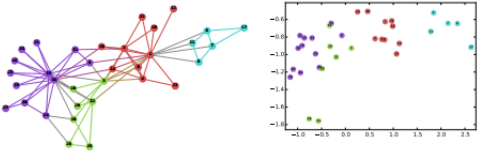
RW-HDP combines random walks and topic model for community detection.

Conducting random walks is a way of aggregating information. Each random walker is an agent who explores a local part of the network. Combining the information they collected properly, we get a big picture of the network.

Topic models are generative models that generally used for documents analysis.

Deepwalk

1. Conduct short random walks on the network
2. Treat them as sentences
3. Deep learning for word (vetex) embedding
4. Use embedding for other tasks



(a) Input: Karate Graph (b) Output: Representation

SSN-LDA

In this model, each node is associated with a social interaction profile, which only takes a node's immediate and secondary neighbors into consideration. Those social interaction profiles are treated as documents for community detection using Latent Dirichlet Allocation.

Basic idea

1. Conduct random walks on the network. Treat vertexes as words, communities as topics, and random walks as documents
2. Use topic model to find the topic structure of each document
3. Use Bayes theorem to find the probabilities of a vertex belonging to different topics and classify it to the largest one

Random walks

Let N_d be the length of the d^{th} random walks. The length can either be fixed or is a Poisson random variable.

Randomly sample D vertexes from the network as starting points of random walks.

Conduct D random walks and treat them as documents.

HDP topic model: stick breaking construction

1. Draw an infinite number of topics, $\beta_k \sim Dirichlet(\eta), k = 1, 2, \ldots$
2. Draw corpus breaking proportions, $v_k \sim Beta(1, \gamma), k = 1, 2, \ldots$
3. For each document:
    1. Draw document-level topic indexes, $c_{di} \sim Multinomial(\sigma(v)), i = 1, 2, \ldots$
    2. Draw document breaking proportions, $\pi_{di} \sim Beta(1, \alpha), i = 1, 2, \ldots$
    3. For each word:
        1. Draw topic assignment $z_{dn} \sim Multinomial(\sigma(\pi_d))$.
        2. Draw word $w_{dn} \sim Multinomial(\phi_{c_{d,z_{dn}}})$.

Graph representation

There should be a picture but I don't have time.

Conditional distributions

Using Markov blanket we get the following conditional distributions

$$p(z_{dn}^i = 1 | \pi_d, \beta_{1:\infty}, w_{dn}, c_d) \propto \exp\{\log \sigma_i(\pi_d) + \sum_{k=1}^{\infty} c_{di}^k \log \beta_{k,w_{dn}}\}$$

$$p(c_{di}^k = 1 | \nu, \beta_{1:\infty}, w_d, z_d) \propto \exp\{\log \sigma_k(\nu) + \sum_{n=1}^{N} \log \beta_{k,w_{dn}}\}$$

$$p(\pi_{di} | z_d) \sim \text{ Beta}(1 + \sum_{n=1}^{N} z_{dn}^i, \alpha + \sum_{n=1}^{N} \sum_{j>i} z_{dn}^j)$$

$$p(v_k | c) \sim \text{ Beta}(1 + \sum_{d=1}^{D} \sum_{i=1}^{\infty} c_{di}^k, \omega + \sum_{d=1}^{D} \sum_{i=1}^{\infty} \sum_{j>k} c_{di}^j)$$

$$p(\beta_k | z, c, w) \sim \text{ Dirichlet}(\eta + \sum_{d=1}^{D} \sum_{i=1}^{\infty} c_{di}^k \sum_{n=1}^{N} z_{dn}^i w_{dn}).$$

Notice that all of them are in Exponential families.

Stochastic variational inference

Based on the conditional distributions, we using the following variational family under the mean field assumption

$$
\begin{aligned}
q(\beta, \nu, z, \pi) = & \Big(\prod_{k=1}^{K} q(\beta_k | \lambda_k) q(\nu_k | a_k) \Big) \times \\
& \Big(\prod_{d=1}^{D} \prod_{i=1}^{T} q(c_{di} | \zeta_{di}) q(\pi_{di} | \gamma_{di}) \prod_{n=1}^{N} q(z_{dn} | \phi_{dn}) \Big)
\end{aligned}
$$

The latent variables c_{di}, π_{di}, z_{dn} depends only on a single document, while the global variables v_k, β_k depends on all documents. To efficiently update the proxy, we can use Stochastic Variational Inference method.

Bayes theorem for community assignment

...

data sets

1. yeast: a yeast protein complex interaction network (Yu *et al* 2008).
2. GSE: a breast cancer gene co-expression network (Chen *et al* 2010, Chen and Xu 2005).
3. ca-GrQc: Arxiv General Relativity and Quantum Cosmology collaboration network. If an author i co-authored a paper with author j, the graph contains an undirected edge between i and j (Leskovec *et al* 2007).
4. ca-CondMat: Arxiv Condense Matter Physics collaboration network (Leskovec *et al* 2007).
5. US powergrid: the high-voltage power grid in the Western States of the United States of America. The nodes are transformers, substations, and generators, and the ties are high-voltage transmission lines (Watts *et al* 1998).

data sets

Table: Network Statistics

statistics	yeast	GSE	ca-GrQc	ca-CondMat	US powergrid
type	biology	biology	co-authorship	co-authorship	engineer
nodes	1540	9112	5242	16264	4941
edges	8703	244928	14478	47594	6594

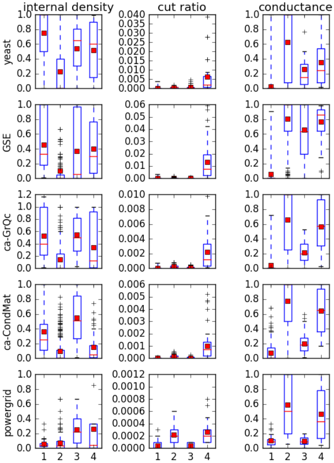
evaluation metrics

1. Internal density: $D = \frac{2m_S}{n_S(n_S-1)}$. This metric sores the community structure based on its internal connectivity. A larger internal density usually means a better community structure (Radicchi *et al* 2004).
2. Cut Ratio: $CR = \frac{c_S}{n_S(n-n_S)}$, which quantifies the community structure based on its external connectivity. A smaller cut ratio usually means a better community structure (Fortunato 2010).
3. Conductance: $C = \frac{c_S}{2m_S+c_S}$, which measures the fraction of edge that points outside the cluster. It combines both internal and external connectivity to give a score. A smaller conductance usually means a better community structure (Shi and Malik 2000).
4. Modularity: $Q = \sum_{i=1}^{m}(e_{ii} - a_i^2)$, where m is the number of communities, e_{ij} 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 $[-0.5, 1)$. A larger modularity means a better community structure (Newman 2006).

comparison models

1. SSN-LDA (Zhang *et al* 2007), a topic based community detection model.
2. Walktrap (Pons and Latapy 2006), a random walk based community detection model. This method does not actually implement random walks on the network, but it defines node-to-node distance and community-to-community distance based on properties of random walks, such as the transition probability between any pair of nodes within t steps. Later, it merges communities iteratively to get a hierarchical tree of partition. Finally, it cuts the tree to get the best partition.
3. BCD (Morup and Schmidt 2012), a nonparametric Bayesian network generative model. The generative process is: first, a cluster assignment is generated using Chinese Restaurant Process (a commonly used metaphor for Dirichlet Process); then, within-cluster and between-cluster link probabilities are generated; finally, links between nodes are generated according to the within- and between-cluster link probabilities.

results



internal density
cut ratio
conductance
yeast
GSE
ca-GrQc
ca-CondMat
powergrid
1
2
3
4

results

Table: Modularity

model	yeast	GSE	ca-GrQc	ca-CondMat	US powergrid
RW-HDP	0.7605	0.5967	0.7848	0.7588	0.9087
SIP2-LDA	0.6995	0.5881	0.7479	0.6615	0.7775
Walktrap	0.6968	0.6014	0.7430	0.7238	0.8953
BCD	0.6452	0.2017	0.5378	0.5041	0.4802

Table: Perplexity

model	yeast	GSE	ca-GrQc	ca-CondMat	US powergrid
RW-HDP	62.26	1124.51	504.16	1262.18	235.46
SIP2-LDA	279.95	1664.80	2902.81	41920.72	7197.49

Pros and Cons

Pros

1. Nonparametric topic model allow community number auto detection
2. Soft-clustering
3. High accuracy compared to other generative models
4. Can be extended to online setting

Cons

1. The inference of probabilistic model is always slow, even SVI is used

Future works

1. Include teleportation to allow single agent to explore a larger area of the network
2. Hyperparameters tuning
3. Ground truth benchmarks comparison

References

...

Thank You