Unsupervised Machine Learning and Data Mining (DS 5230)
Spring 2022 – Project Posters – 100 points
Due: Wednesday, May 4, 2022 at 11:59 PM Eastern
(There are no late days for this assignment.)

For your project poster, **use the template at https://tinyurl.com/yxjyfx3v** to prepare a poster (in either Powerpoint or PDF) and upload it to Canvas by Wednesday, May 4, 2022 at 11:59 PM Eastern. Your file should be named

<Lastname><FirstName>ProjectPoster.pdf

or

<Lastname><FirstName>ProjectPoster.pptx.

To make the grading process easier, I would like you to upload your presentation -- even if you are part of a team.

Your poster should have the following in it:

- Title & Author(s)
- Problem Definition: be concrete as possible
- Related Work / Existing Methods
- Proposed Methods / Algorithms
    - Give the overview of the approach.
- Data Description & Experimental Setup
- Key Results & Discussion
    - Present key results and contributions.
    - Do not superficially cover all results; cover key result well.
    - Do not just present numbers; interpret them to give insights (in terms of contributions).
- Takeaway Points & Future Work
    - Give problems this research opens up (optional).

Again, use the template at https://tinyurl.com/yxjyfx3v. Below you will find a sample project poster. A PDF version of the sample project poster is available at http://eliassi.org/sample-poster.pdf.

HCDF: Hybrid Community Discovery Framework

Tina Eliassi-Rad*
tina@eliassi.org

Problem Definition

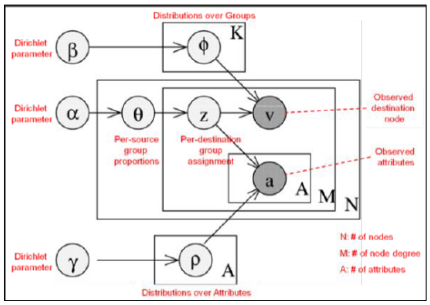
- Given a **graph** G=(V, E), detect **communities** of vertices
- Want a community discovery procedure with the following properties
  - **Scalable**, where time and space complexity are strictly sub-quadratic w.r.t. the number of nodes
  - **Nonparametric**, where number of communities need not be specified a priori
  - **Consistent**, where effectiveness is consistently high across a wide range of domains
  - **Effective**, where global connectivity patterns are successfully factored into communities that are highly predictive of individual links and robust to small perturbations in network structure

Existing Methods

- Hard Clustering: each vertex belongs to a single community
  - Fast Modularity (FM) [Clauset+, Phys. Rev. E. 2004]
    - Prefers communities with high intra-community connectivity
  - Cross-Associations (XA) [Chakrabarti+, ACM KDD 2004]
    - Prefers communities where nodes in a given row-group or column-group have similar connectivity to each other
    - Minimizes total encoding cost of adjacency matrix
- Soft Clustering: vertices allowed mixed membership
  - Latent Dirichlet Allocation for Graphs (LDA-G) [Henderson & Eliassi-Rad, ACM SAC 2009]
    - Application of Latent Dirichlet Allocation topic model to graphs
    - Learned "topics" or communities maximize likelihood of observed edges

Proposed Method

- LDA-G used as core Bayesian model for community discovery
- A hard clustering method is applied to the graph
- Resulting communities are used as hints for the Bayesian model
  - There exist multiple strategies for incorporating hints
  - The most effective is to add hard cluster labels as attributes and extend the Bayesian model
- HCD (or HCD-X) = LDA-G with XA communities as attributes
- HCD-M = LDA-G with FM communities as attributes
- Algorithm:
  - Run XA (or FM) on input G=(*V*, *E*)
    - Produces groups, *A*, over nodes
  - Run LDA-G on graph *G'* = (*V*, *E*, *A*)
- Use Gibbs sampling to infer posterior estimates on *group* and *source-node* distributions



See report for computational complexity discussion.

Data Description & Experimental Setup

Real-World Graphs	Acronym	| *V* |	| *E* |
Autonomous Systems Graph	AS	11,461	32,730
Day 1: IP × IP	IP1	34,449	303,175
Day 2: IP × IP	IP2	33,732	320,754
Day 3: IP × IP	IP3	34,661	428,596
Day 4: IP × IP	IP4	34,730	425,368
Day 5: IP × IP	IP5	33,981	112,271
PubMed Author × Knowledge	AxK	37,436 (A) 117 (K)	119,443
PubMed Coauthorship	AxA	37,227	143,364
WWW Graph	WWW	325,729	1,497,135

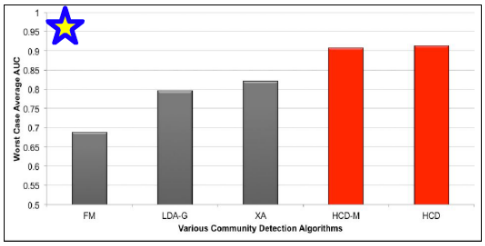
Link prediction experimental setup

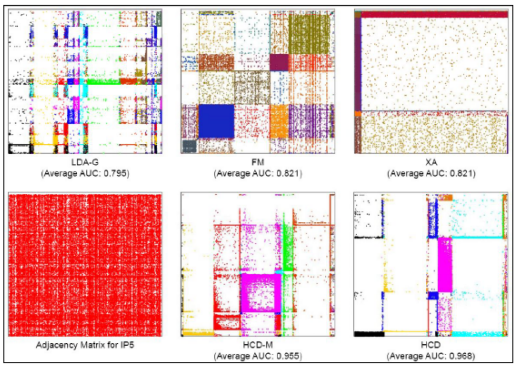
- For i=1 to 5 do
  - Hold out 500 "present" edges and 500 "absent" edges
  - Run model on remaining graph
  - Model scores held-out links
  - Compute Area Under ROC curve (AUC)
- Report average & standard deviation of AUCs
- High AUC = good link prediction

For robustness experiments, see final report.

Results & Discussion

- **Consistency:** Across a variety of domains
  - Non-hybrid methods struggle with at least one graph
  - Hybrid methods always perform well
- **Effectiveness:** Hybrid methods never perform significantly worse than their constituents
- **Better compression:** Reordering by discovered communities shows that hybrid methods exhibit better compression (more whitespace) than non-hybrids.
- **Good link prediction is a tradeoff between low entropy & flexibility**
  - Low entropy
    - If the adjacency matrix can be **compressed nicely** or mixed-membership distributions are **far from uniform**, we can better predict behavior of nodes
  - Flexibility
    - If a node exhibits multiple types of behavior, hard clustering may only model **a plurality** of the node's edges, which can explain all links





The better the compression, the higher the AUC.

Takeaway Points & Future Work

- Use a **hybrid approach** to community discovery on graphs for **consistent**, **effective**, **Nonparametric** community factorization on graphs from **various domains**
- Incorporate **hints as attributes** for coalescing strategy
- Use **link prediction** and **variation of information** as a **quantitative measure** on the communities discovered
- **Future work:** Extension to time-evolving graphs

* Joint work with Keith-Henderson, Spiros Papadimitriou, & Christos Faloutsos. Report at http://eliassi.org/papers/henderson-sdm10.pdf.