INDY EPFL

# Network Community Structure under Metric Sparsification


Maximilien Dreveton, Charbel Churcri, Matthias Grossglauser, Patrick Thiran

INDY (Information and Network Dynamics Lab)

School of Communication and Computer Sciences

EPFL

CH-1015 Lausanne

Patrick.Thiran@epfl.ch

http://indy.epfl.ch

# Graph Sparsification

- Weighted graph $G(V,E,c)$. NB: $|E|$ is often O($n$^2), with $n = |V|$.
- Here $c_e$ = cost of edge $e$, sampled from distribution $F$.
- Sparsification of $G$ : prune $E$ so that pruned graph $G^s$ keeps the same structural properties.
- Benefits: visualization, reduced computational and storage cost.
- Extends $c_e$ to unweighted graphs by taking inverse of Jaccard index for edge $e = (u,v)$:

$$c_e = \frac{|\mathrm{Nei}(u) \cup \mathrm{Nei}(v)|}{|\mathrm{Nei}(u) \cap \mathrm{Nei}(v)|} - 1$$

- Threshold Sparsification
  - Keep edge $e$ in $G^s$ iff $c_e < c_{th}$ for some threshold $c_{th}$.

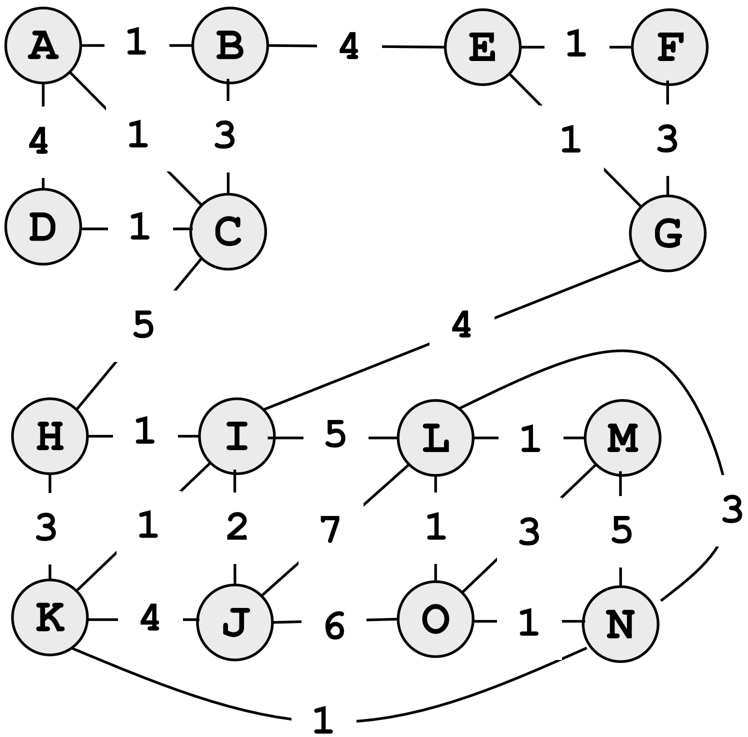
# Graph Sparsification

- Spectral Sparsification [Spielman, Srivtsava, 2011]
  - Keep edge *e* with probability $s_e \sim R_e/c_e$ where $R_e$ = effective resistance of edge *e* (proportional to the probability that edge *e* appears in a random spanning tree of *G*), and reweight $c_e$
  - For $s_e$ large enough, $G^s = G^{ss}$ maintains spectral properties. (Laplacians $L_{G^{ss}} \approx L_G$)

$$(1-\varepsilon)\lambda(L_G) \leq \lambda(L_{G^{ss}}) \leq (1+\varepsilon)\lambda(L_G)$$

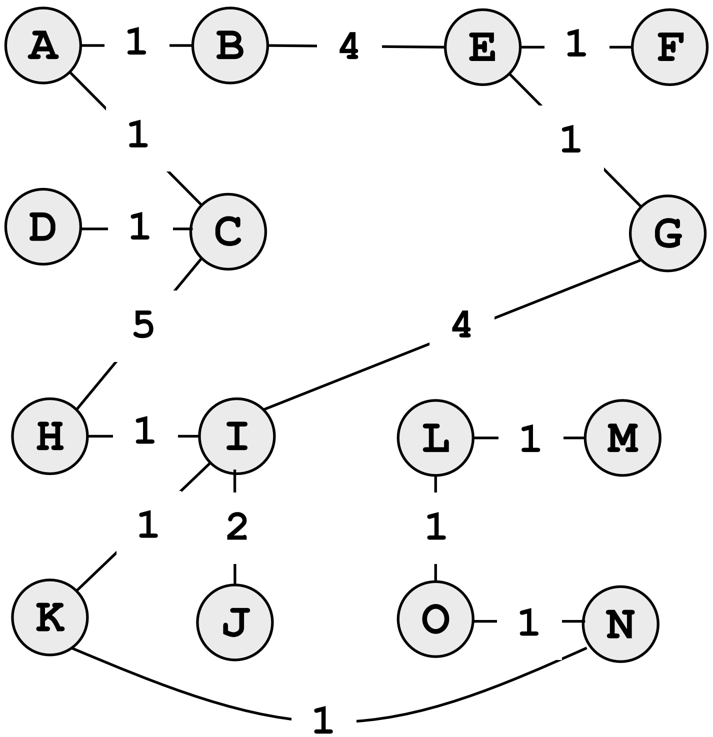
- Metric Sparsification: Metric Backbone $G^s = G^{mb}$
  - Keep edge *e* iff it is metric (appears in some shortest path between 2 vertices in *V*).
  - No hyperparameter.
  - Same Idea behind betweeness Community Detection [GirvanNewman2004]: edges traversed by the highest number of shortest paths separate communities.

# Metric Backbone



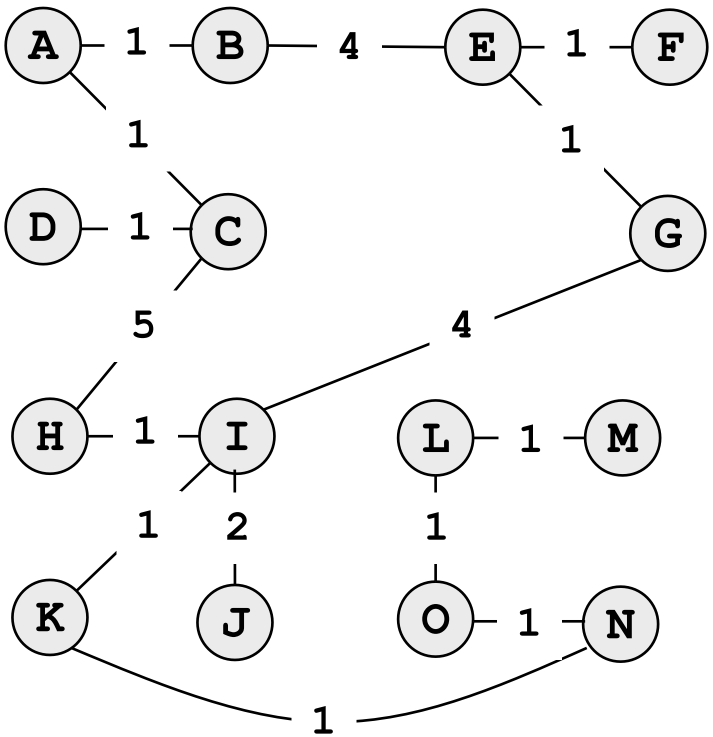

- Metric Sparsification: Metric Backbone $G^s = G^{mb}$
  - Keep edge *e* iff it is metric (appears in some shortest path between 2 vertices in *V*).
  - Maintains shortest path distances, betweenness centrality, pagerank, connected components.

# Metric Backbone




- Metric Sparsification: Metric Backbone $G^s = G^{mb}$
  - Keep edge *e* iff it is metric (appears in some shortest path between 2 vertices in *V*).
  - Maintains shortest path distances, betweenness centrality, pagerank, connected components.

# Metric Backbone



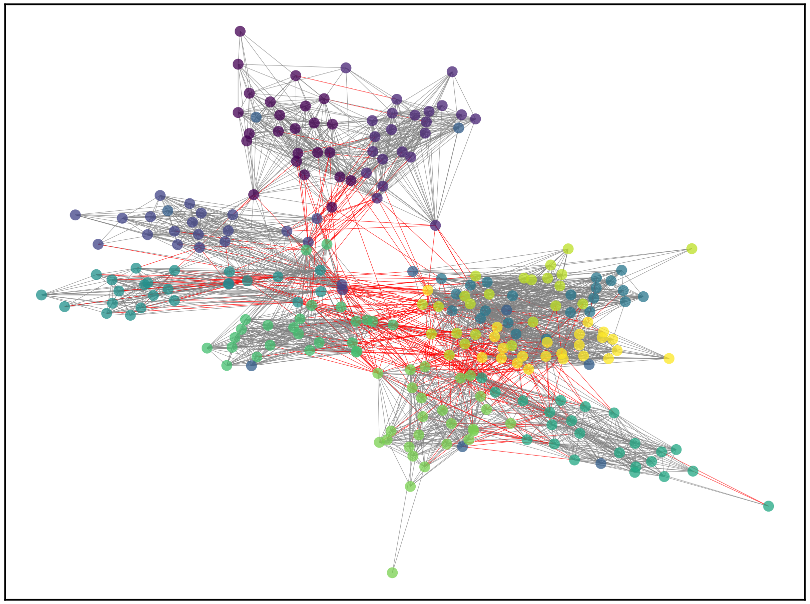

- Metric Sparsification: Metric Backbone $G^s = G^{mb}$
  - Keep edge *e* iff it is metric (appears in some shortest path between 2 vertices in *V*).
  - Maintains shortest path distances, betweenness centrality, pagerank, connected components.
  - Community structure?

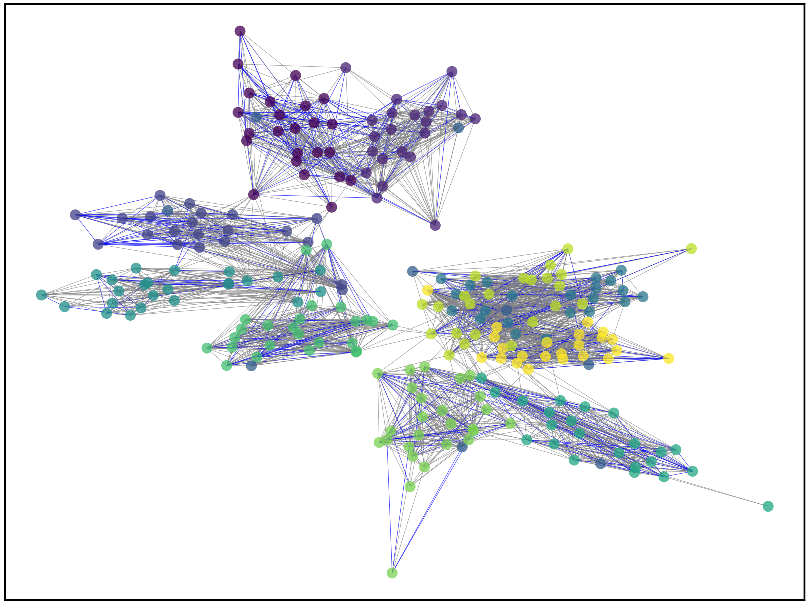
# How Sparse is $G^{mb}$ ?

| Graph | $\|V\|$ | $\|E\|$ | preprocessing | % edges deleted |
|---|---|---|---|---|
| Facebook | 190M | 49.9B | Custom | 26.5% |
| Twitter | 40M | 1.5B | Jaccard | 39% |
| Tuenti | 12M | 685M | Jaccard | 59% |
| LiveJournal | 4.8M | 34M | Jaccard | 40% |
| NotreDame | 0.3M | 1.5M | Jaccard | 45% |
| | | | Adamic | 29% |
| DBLP | 318K | 1M | Jaccard | 23% |
| | | | Adamic | **9%** |
| Twitter-ego | 1.7M | 1M | Jaccard | 57% |
| | | | Adamic | 39% |
| Movielens | 1.6K | 1.9M | Jaccard | **88%** |
| Facebook | 1K | 143K | #messages | 78% |
| US-airports | 30.5K | 6K | #passengers | 72% |
| C-Elegans | 0.3K | 2.3K | #connections | 17% |

From Kalavri, Simas, Logothetis (2016). The shortest path is not always a straight line: leveraging semi-metricity in graph analysis. Proceedings of the VLDB Endowment, 9(9), 672-683.

# Metric Backbone vs Threshold Graph

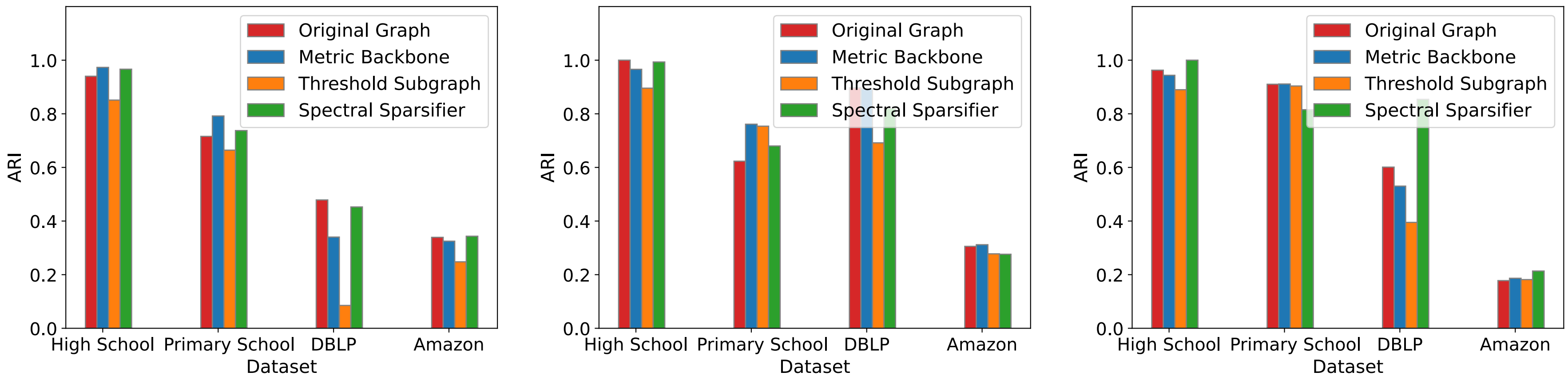




- ❑ Primary school dataset, threshold set to keep same % of edges
- ❑ Empirical Evidence that Metric Backbone preserves Community Structure.

# Graph Sparsification and Clustering

- Effect of sparsification on performance of clustering algorithms




**(a)** *Bayesian MCMC* **(b)** *Leiden* **(c)** *Spectral Clustering*

- Extends observations in Brattig Correia, R., Barrat, A., Rocha, L. M. (2023). Contact networks have small metric backbones that maintain community structure and are primary transmission subgraphs. PLoS Computational Biology, 19(2), e1010854.

# Weighted Stochastic Block Model

- *n* nodes in *k* latent blocks.
- $z_u$ = block membership of node *u* (i.i.d) $\mathbb{P}(z_u = a) = \pi_a,\ a \in \{1, 2, \ldots k\}$
- $p_{ab}$ = Prob(edge between a node in block *a* and a node in block *b*)
- *c(u.v)* = cost of edge *(u,v),* sampled from cdf $F_{ab}$.
- $(z, G) \sim \mathrm{wSBM}(n, \pi, p, F)$ and $G = ([n], E, c),\ z \in [k]^n$

$$
\begin{aligned}
\mathbb{P}(z) &= \prod_{u=1}^{n} \pi_{z_u}, \\
\mathbb{P}(E \mid z) &= \prod_{1 \leq u < v \leq n} p_{z_u z_v}^{1\{(uv) \in E\}} \left(1 - p_{z_u z_v}\right)^{1\{(uv) \notin E\}}, \\
\mathbb{P}(c \mid E, z) &= \prod_{(u,v) \in E} \mathbb{P}\left(c(u,v) \mid z_u, z_v\right).
\end{aligned}
$$

# Assumptions

- Asymptotic scaling: $p_{ab} = B_{ab}\rho_n$ with
  - $\rho_n = \omega(\log n/n)$
  - For all $a, b \in [k]$: $B_{ab} > 0,\ \pi_a > 0$ fixed.
- Costs sampled from fixed cdf $F_{ab}$
  - Continuous
  - For all $a, b \in [k],\ F_{ab}(0) = 0,\ F'_{ab}(0) := \lambda_{ab} > 0$
- Some notations:

$$\tau_{\max} := \max_{a\in[k]} \sum_{b=1}^{k} \lambda_{ab} B_{ab} \pi_b$$

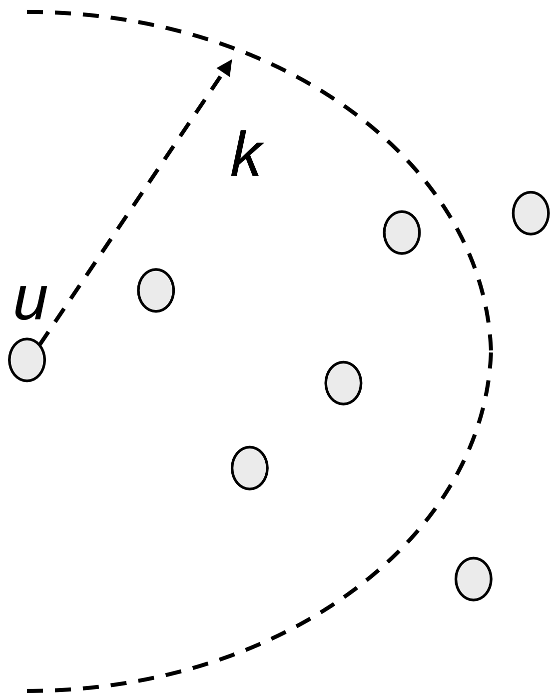
$$\tau_{\min} := \min_{a\in[k]} \sum_{b=1}^{k} \lambda_{ab} B_{ab} \pi_b$$

# Cost of shortest paths on wSBM

- Cost C(u,v) of shortest path between any pair of nodes *u,v* chosen uniformly at random in their blocks is whp

$$\frac{1}{\tau_{\max}} \leq \frac{n\rho_n}{\log n} C(u,v) \leq \frac{1}{\tau_{\min}}$$

- Proof for $F_{ab}$ ~ expo($\lambda_{ab}$)
  - First Passage Percolation (FPP) from *u* until their $k^{th}$-nearest neighbors




  - Let $C_u(k)$ be the cost to $k^{th}$-nearest neighbor of *u*
  - Conditioned on edges exposed from the previous *k* neighbors from *u*,

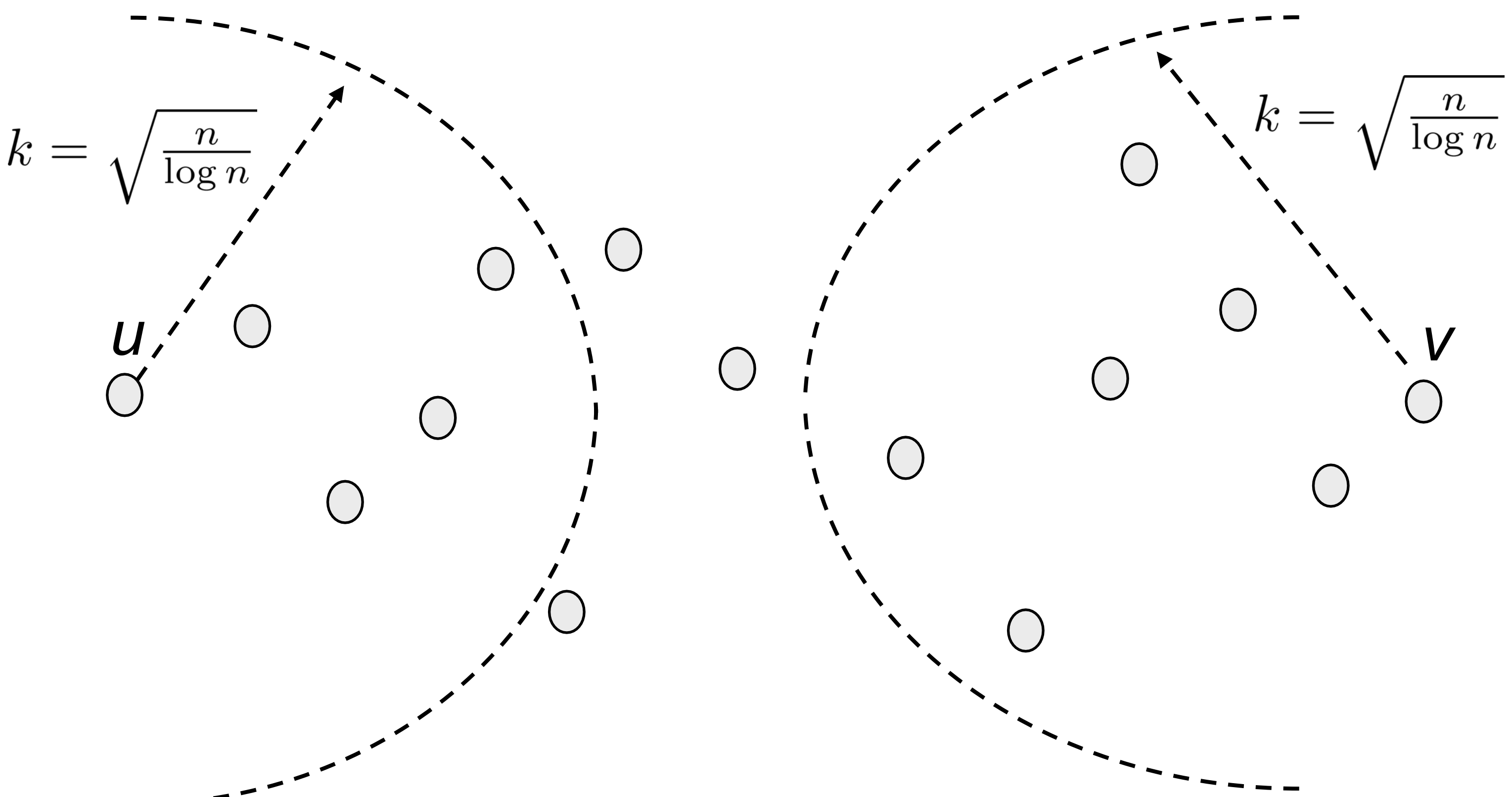
    $C_u(k+1) - C_u(k)$ ~ iid expo.
  - Can compute that whp

$$C_u(\sqrt{n \log n}) \leq \frac{1+o(1)}{2\tau_{\min}} \frac{n\rho_n}{\log n}$$

$$C_u\left(\sqrt{\frac{n}{\log n}}\right) \geq \frac{1+o(1)}{2\tau_{\max}} \frac{n\rho_n}{\log n}$$

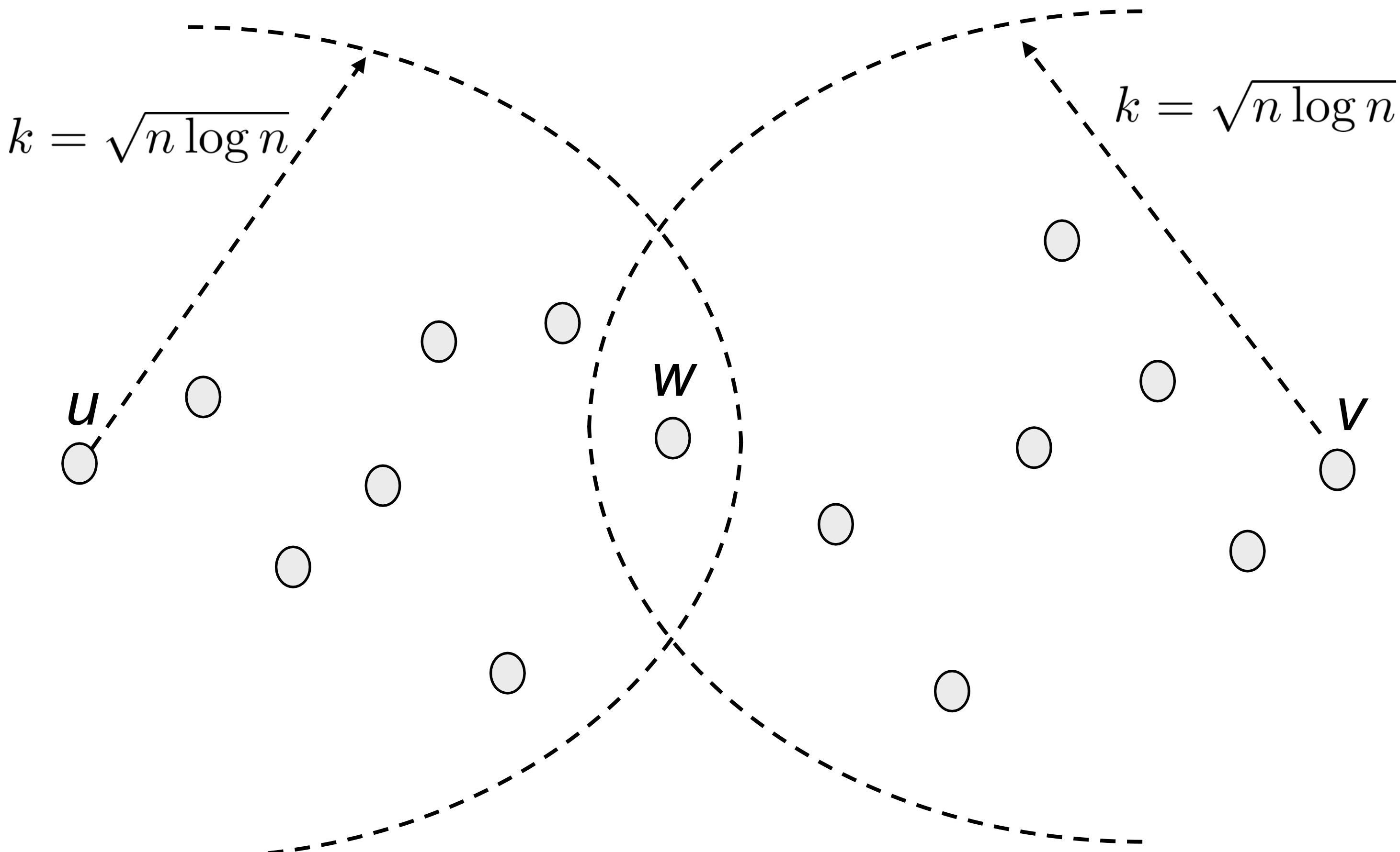
# Cost of shortest paths on wSBM

- Two FPPs from *u* and *v*
- Case 1: $k = \sqrt{\frac{n}{\log n}} \Longrightarrow C_u\left(\sqrt{\frac{n}{\log n}}\right) \geq \frac{1+o(1)}{2\tau_{\max}} \frac{n\rho_n}{\log n}$
  - FPPs from u and v have an empty intersection whp
  - $C(u,v) \geq C_u\left(\sqrt{\frac{n}{\log n}}\right) + C_v\left(\sqrt{\frac{n}{\log n}}\right) \geq \frac{1+o(1)}{\tau_{\max}} \frac{n\rho_n}{\log n}$



# Cost of shortest paths on wSBM

- Case 2: $k = \sqrt{n \log n} \implies C_u(\sqrt{n \log n}) \leq \frac{1+o(1)}{2\tau_{\min}} \frac{n\rho_n}{\log n}$
  - FPPs from *u* and *v* have a non-empty intersection whp
  - There is some $w \in C_u(\sqrt{n \log n}) \cap C_v(\sqrt{n \log n})$
  - $C(u,v) \leq C(u,w) + C(w,v) \leq \frac{1+o(1)}{\tau_{\min}} \frac{n\rho_n}{\log n}$



# Cost of shortest paths on wSBM

- Cost $C(u,v)$ of shortest path between any pair of nodes $u,v$ chosen uniformly at random in their blocks is whp

$$\frac{1}{\tau_{\max}} \leq \frac{n\rho_n}{\log n} C(u,v) \leq \frac{1}{\tau_{\min}}$$

- Proof for $F_{ab}$ continuous such that

$$\forall a, b \in [k],\ F_{ab}(0) = 0,\ F'_{ab}(0) := \lambda_{ab} > 0$$

  - Argument from [Janson, 1999]. One, two and three times log n/n for paths in a complete graph with random weights.
  - $G = (V,E,c) \sim \text{wSBM}(n,p,\pi,(F_{ab})_{ab})$
    $\rightarrow G_{\text{unif}} = (V,E,c_{unif}) \sim \text{wSBM}(n,p,\pi,(\text{unif}(0,1/\lambda_{ab}))_{ab})$
    $\rightarrow G_{\text{exp}} = (V,E,c_{\text{exp}}) \sim \text{wSBM}(n,p,\pi,(\text{expo}(\lambda_{ab}))_{ab})$
  - Show that the shortest paths in $G_{\text{exp}}$, $G_{\text{unif}}$ and $G$ are the same.

# Cost of shortest paths on wSBM

- Cost *C(u,v)* of shortest path between any pair of nodes *u,v* chosen uniformly at random in their blocks is whp

$$\frac{1}{\tau_{\max}} \leq \frac{n\rho_n}{\log n} C(u,v) \leq \frac{1}{\tau_{\min}}$$

- Corollary: probability of keeping edge in Metric Backbone of wSBM.
  - Remember $p_{ab} = \mathbb{P}((u,v) \in E \mid z_u = a, z_v = b)$
  - Let $p_{ab}^{mb} = \mathbb{P}((u,v) \in E^{mb} \mid z_u = a, z_v = b)$
  - Then for any pair of nodes *u,v* chosen uniformly at random in blocks *a* and *b*, whp

$$(1+o(1))\frac{1}{\tau_{\max}} \leq \frac{n\rho_n}{\log n} \frac{p_{ab}^{mb}}{p_{ab}} \leq (1+o(1))\frac{1}{\tau_{\min}}$$

- Proof: adapt [Corollary 1, vanMieghemW, 2009]

# Example: Planted Partition Model

- For any pair of nodes *u, v* chosen uniformly at random in blocks *a, b*, whp

$$(1+o(1))\frac{1}{\tau_{\max}} \leq \frac{n\rho_n}{\log n}\frac{p_{ab}^{mb}}{p_{ab}} \leq (1+o(1))\frac{1}{\tau_{\min}}$$

- Let $p_{ab} = B_{ab}\rho_n$ with $\rho_n = \omega(\log n/n)$ and
  - For all $a, b \in [k]$: $\pi_a = \frac{1}{k}$ , $\lambda_{ab} = \lambda > 0$ and $B_{ab} = \begin{cases} p_0 & \text{if } a = b, \\ q_0 & \text{otherwise} \end{cases}$
  - Then $\tau_{\max} = \tau_{\min} = \lambda(p_0 + (k-1)q_0)/k$ and

$$\begin{aligned} p^{mb} &= (1+o(1))\frac{kp_0}{p_0+(k-1)q_0}\frac{\log n}{n} \\ q^{mb} &= (1+o(1))\frac{kq_0}{p_0+(k-1)q_0}\frac{\log n}{n} \end{aligned}$$

  - In particular,

$$\frac{p^{mb}}{q^{mb}} = (1+o(1))\frac{p_0}{q_0}$$

# Spectral Clustering on the Metric Backbone

- Algorithm
  - Input: Graph *G*, number of clusters *k*.
  - Output: Predicted community memberships $\hat{z}_v, v \in V$
  - *W* = weighted adjacency matrix of *G*, with eigendecomposition

$$W = \sum_{i=1}^{n} \sigma_i u_i u_i^T = U \Sigma U^T, \Sigma = \mathrm{diag}(\sigma_1, \cdots, \sigma_k), U = [u_1, \cdots, u_k]$$

  - $\hat{z}$ = (1+ε)-approximate solution of k-means performed on rows of *U*
- Loss for *G* = $G^{mb}$, with Sym(*k*) = set of all permutations of [*k*]:

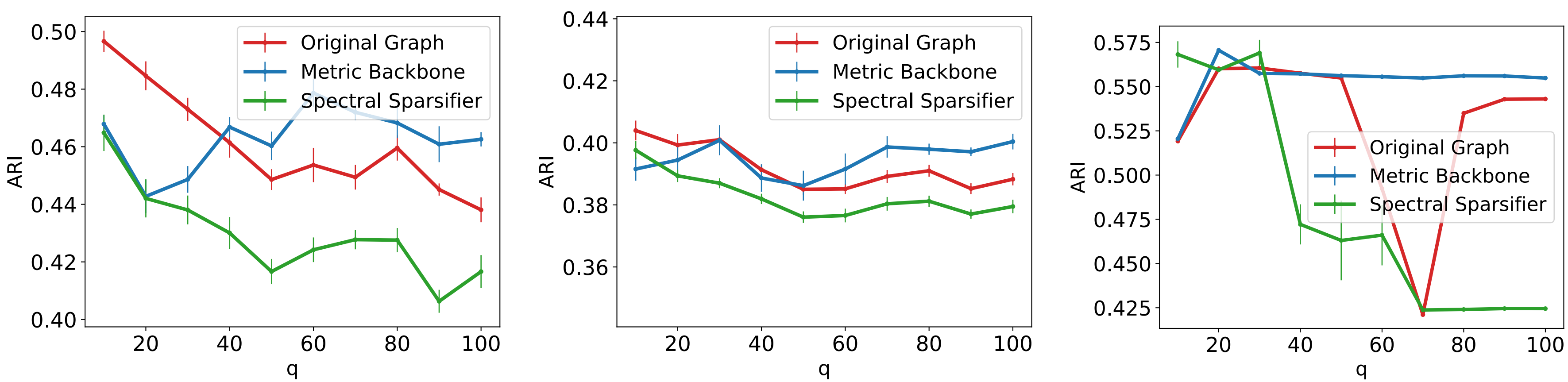
$$\mathrm{loss}(z, \hat{z}) = \frac{1}{n} \inf_{\tau \in \mathrm{Sym}(k)} \text{Hamming Distance}\left(z, \tau \circ \hat{z}\right)$$

- If $\tau_{\max} = \tau_{\min}$ and $\mu$ = minimal eigenvalue of $[\lambda_{ab} B_{ab} \pi_b]_{a,b}$ is non zero, then whp

$$\mathrm{loss}(z, \hat{z}) = O\left(\frac{1}{\mu^2 \log n}\right)$$

# Graph Construction

- Construct proximity graph $G = ([n], E, p)$ from $n$ data points $x_1, \ldots, x_n \in R^d$.
  - similarity function sim($x_u$,$x_v$)
  - $N(u,q)$ = {$q$ items the most similar to $u$);
  - proximity associated with edge(u,v) = $(s_{uv} + s_{vu})/2$ where $s_{uv}$ = sim($x_u$,$x_v$) if $v \in N(u,q)$ and 0 otherwise.
- Common choice: Gaussian kernel similarity sim($x_u$,$x_v$) ~ $\exp(-\|x_u - x_v\|^2)$




**(a)** *MNIST* **(b)** *Fashion MNIST* **(c)** *HAR*

# Conclusion

- Communities are well-preserved by the metric backbone.
- Theoretical confirmation on wSBMs using FPP techniques [KolossvaryK, 2015].
- Shortest paths of wSBMs are longer than shortest paths in real networks (hop-count $\Theta(\log n)$ in wSBMs $\Theta(1)$ in real networks).
- Extension to unweighted networks.