Centro Interdisciplinario en Ciencia de Datos y Aprendizaje Automático

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Embeddings Hiperbólicos de Grafos

Paola Bermolen

Seminario Optimización y Machine Learning
Junio 2025

Proyecto: Geometria de Redes Complejas y Aplicaciones al Aprendizaje Automático (CSIC I+D)

Bernardo Marenco

Marcelo Fiori

Federico Larroca

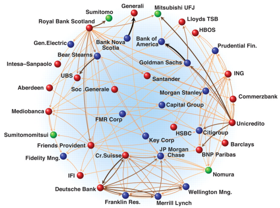
Gonzalo Mateos

Sofia Perez

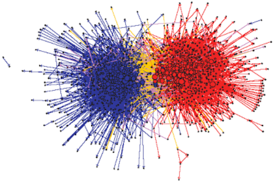
Matias Carrasco

Real Networks & Graphs

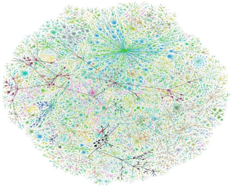
Economic Networks



Social and Information Networks

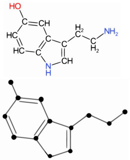


Internet

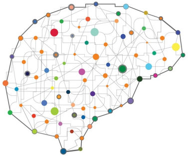


3D Meshes

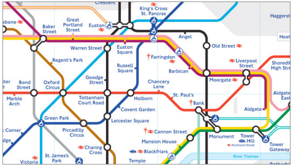
Molecules



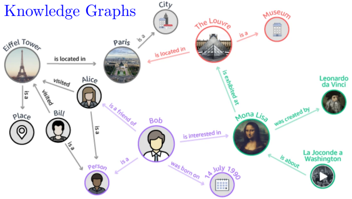
Brain Connectomes



Transportation Networks



Knowledge Graphs



Algunas propiedades que se repiten: distribución de grados power-law, sparse, diametro pequeño, estructura de comunidades y clustering alto... no facilmente reproducibles con modelos clásicos de grafos aleatorios

Motivación

- Venimos estudiando el problema de "Graph Representation Learning" (GRL) en general y el modelo Random Dot Product Graphs (RDPG) en particular

- ¿Qué es GRL? busco representación de menor dimensión (embedding) que capture las propiedades más importantes del grafo

  $\Rightarrow$ distancia o similitud entre los embeddings refleje similitud "semántica" entre los nodos

- Algunas limitantes: las redes reales no son tan RDPGs y se pueden precisar dimensiones muy altas para lograr representaciones adecuadas

- Pregunta: ¿si usamos embeddings hiperbólicos podemos capturar mejor las propiedades del grafo y con dimensiones más chicas?

Contexto

- $\approx$ 2010, desde la fisica (mecánica estadistica) y el análisis de redes/sistemas complejos se trabaja sobre la idea de que las redes complejas reales tienen una geometria subyacente hiperbólica
- Algunas ideas clave:

  $\Rightarrow$ posicionar los nodos en un disco hiperbólico y asignar aristas con probabilidad decreciente con la distancia entre los nodos produce naturalmente distribuciones de grado power law y alto niveles de clustering...

  $\Rightarrow$ los espacios hiperbólicos son versiones continuas de árboles y las redes complejas son localmente árboles

Contexto

- $\approx$ 2017, explotó la comunidad de ML con métodos hiperbólicos (Standford, MILA, ETH, Facebook...)..

  $\Rightarrow$ varios trabajos muestran desempeños similares o mejores con dimensiones más bajas

  $\Rightarrow$ mucha falta de rigurosidad, de buenos benchmark y ni hablar de garantiaas matemáticas

- Nosotros:
  1. Tratar de tener criterios y fundamentos teóricos para decidir cuándo es mejor usar embeddings hiperbólicos
  2. Extender a escenarios dinámicos: aprovechando el enfoque de optimización en variedades
  3. Armar una plataforma única con los principales algoritmos para poder comparar

Ideas de los fisicos

Physical Review E

Highlights Recent Accepted Collections Authors Referees Press About Editorial Team RSS

Hyperbolic geometry of complex networks

Dmitri Krioukov[1], Fragkiskos Papadopoulos[2], Maksim Kitsak[1], Amin Vahdat[3], and Marián Boguñá[4]

Show more

Phys. Rev. E **82**, 036106 – **Published 9 September, 2010**

DOI: https://doi.org/10.1103/PhysRevE.82.036106

....most famous one is general relativity, interpreting gravitation as a curved geometry.
.. yet another example is the recent conjecture by Palmer suggesting that many mysteries of quantum mechanics can be resolved by the assumption that a hidden fractal geometry underlies the universe.....

Espacio Hiperbólico y Árboles

- En un árbol con factor de ramificación constante b se tiene:
  - Cantidad de nodos a distancia r de la raiz: $(b+1)b^{r-1}$
  - Cantidad de nodos a distancia $\leq r$ de la raiz: $\frac{(b+1)b^r-2}{b-1}$

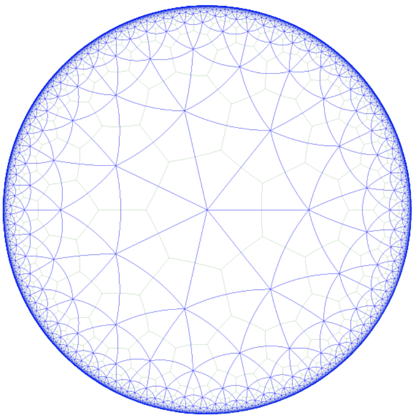
    $\Rightarrow$ Crece como b^r, esto es exponencialmente con r

- Sea $\mathbb{H}^2_\zeta$ plano hiperbólico de curvatura constante $-\zeta^2 < 0$
  - Largo del circulo de radio r: $L(r) = 2\pi \sinh \zeta r$
  - Área del circulo de radio r: $A(r) = 2\pi(\cosh \zeta r - 1)$

    $\Rightarrow$ Crece como $\exp \zeta r$
- Si $\zeta = \ln(b)$, la estructura métrica de $\mathbb{H}^2_\zeta$ y de los b-árboles son similares.
- Para los árboles (incluso infinitos) podemos definir embbedings hiperbólicos casi isométricos

  $\Rightarrow$ se necesita un espacio de tamaño (euclideo) exponencial para la ramificación

Espacio Hiperbólico y Árboles

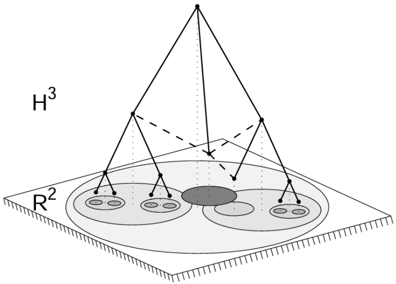


(b)

Figure: Triángulos y heptógonos son de igual tamaño hiperbólico pero el tamaño euclideo decrece exponencialmente con la distancia al centro mientras que la cantidad crece

Topological heterogeneity vs Geometrical Hyperbolicity

- Las redes complejas conectan elementos (nodos) distinguibles y heterógeneos
  $\Rightarrow$ hay alguna taxonomia respecto a la cual se pueden clasificar



- un punto de $\mathbb{R}^2$ representa un atributo y un disco representa un conjunto de atributos para un nodo
- discos superpuestos refieren a nodos con atributos similares
- la coordenada z es el radio del disco: hay linea sólida si uno de los discos es el más chico que contiene al otro
- "casi" árbol

- La distancia entre nodos basadas en la similaridad de atributos se puede mapear a distancias en el espacio hiperbólico ($\mathbb{H}^3$):
  $\Rightarrow$ a más superposición (similitud), menor distancia hiperbólica entre nodos

Geometrias Hiperbólicas producen Topologias Heterogéneas

- Escenario bien simple:

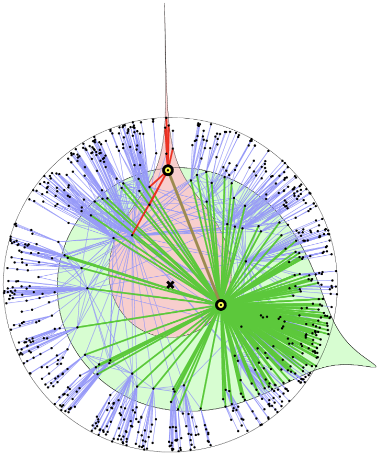
  $\Rightarrow$ Nodos: N puntos uniformes es un disco de radio R (coordenadas polares: $\rho(\theta) = \frac{1}{2\pi}$ y $\rho(r) \approx e^{r-R}$

  $\Rightarrow$ Aristas: si están a menos de R en distancia hiperbólica
- ¿Qué topologia de red aparece en este escenario?
- Grado medio de nodos a distancia r del origen $\hat{k}(r) = \delta A(r)$ de donde $\hat{k} = \int_0^R \rho(r)\hat{k}(r)dr \approx \frac{8}{\pi} N e^{-R/2}$

  $\Rightarrow$ definiendo el grado medio se define el radio R
- Distribución de grados $p(k) \sim k^{-3}$ es power-law

  $\Rightarrow$ viene de la combinación de dos exponenciales, densidad de nodos y grado medio
- Vale también si la curvatura es $-K^2$ y la densidad cuasi-uniforme...

Ejemplo



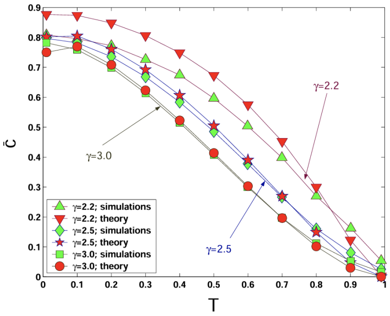
- $N = 740$ nodos, exponente de la power-law $\gamma = 2.2$ y grado medio $\hat{k} = 5$
- disco hiperbólico con $K = -1$ de radio $R = 15.5$.
- visualización (solo ocupa parte pequeña del disco)
- mayoria de conexiones radiales y los nodos periféricos no están conectados entre ellos (dist. hip. gde)
- Discos colores, son discos de radio R a distancia $r = 10.6$ y $r = 5.0$ del origen

Viceversa?

- Prueban también que en una red scale-free con alguna estructura métrica, las distancias métricas ese pueden re-escalar de manera de obtener métricas hiperbólicas

Viceversa?

- Prueban también que en una red scale-free con alguna estructura métrica, las distancias métricas ese pueden re-escalar de manera de obtener métricas hiperbólicas
- Clustering como función de la “temperatura $= \frac{1}{\beta}$”:
  - Coeficiente de clustering promedio se maximiza en $T = 0$ y gradualmente (y casi linealmente) decrece a cero en la transición de fase $T = 1$.



Embeddings hiperbólicos

Para hacer inferencia hay dos familias de métodos:

1. A partir de un modelo generativo donde las aristas están presentes dependiendo de la distancia hiperbólica entre los embeddings según la fórmula:

$$p_{i,j} = [1 + e^{\frac{\beta}{2}(d(\theta_i,\theta_j)-R)}]^{-1}$$

- Métodos: Mercator [2019] (y su extensión multidimensional D-Mercator [2023]), Hyperlink [2020] e HyperMap [2014].
- En los dos primeros, estiman R y β y luego acomodan los puntos en el espacio hiperbólico para que respeten esas distancias.
- HyperMap considera un modelo equivalente de crecimiento del grafo (el denominado Popularity-Similarity-Optimization o PSO [2012]), donde los nodos y sus embeddings se van agregando secuencialmente al grafo.

$\Rightarrow$ Las expresiones matemáticas involucradas para estimar u optimizar son extremadamente complejas y se debe recurrir a aproximaciones o heuristicas.

Embeddings hiperbólicos

Para hacer inferencia hay dos familias de métodos:

2 Asume el conocimiento de una cierta distancia entre los nodos, y busca ubicar los embeddings en el espacio de tal forma de reflejar de la mejor manera estas distancias

- Métodos: Poincaré Maps [2020], Hydra [2020] y Poincaré Embeddings [2017] (o Lorentz Embeddings [2018]).
- Hydra utiliza el largo del camino más corto dos nodos. Transformando las distancias hiperbólicas (cosh) se puede resolver la optimización de manera exacta usando descomposiciones espectrales.
- Poincaré Maps usa el indice denominado Relative Forest Accessibility [?]
- Tanto Poincaré Maps como Poincaré/Lorentz Embeddings realizan descenso por gradiente en la variedad definida por el espacio hiperbólico.

Poincaré Embeddings

NeurIPS Proceedings

Poincaré Embeddings for Learning Hierarchical Representations

Part of Advances in Neural Information Processing Systems 30 (NIPS 2017)

Bibtex Metadata Paper Reviews

Authors

Maximillian Nickel, Douwe Kiela

Poincaré Embeddings

- $S = \{x_i\}_{i=1,\cdots,n}$ conjunto de simbolos
- Poincaré ball model: $(\mathcal{B}^d, g_x)$ con $\mathcal{B}^d = \{x \in \mathbb{R}^d : ||x|| < 1$ y $g_x = (\frac{2}{1-||x||^2})^2 g_E$
- $\Theta = \{\theta_i\}_{i=1,\cdots,n} \subset \mathcal{B}^d$ bola de Poincaré en $\mathbb{R}^d$ de dimensión 1
- Función de pérdida: $\mathcal{L}(\theta)$

$$\hat{\theta} = \underset{\Theta}{\operatorname{argmin}} \, \mathcal{L}(\theta) \quad \text{con} \quad ||\theta_i|| < 1,$$

- Descenso por gradiente en la variedad:

$$\theta_{t+1} = \mathcal{R}_{\theta_t}(-\eta \nabla_R \mathcal{L}(\theta_t))$$

donde

$$\nabla_R \mathcal{L}(\theta_t)) = (\frac{1-||x||^2}{2})^2 \nabla_E \mathcal{L}(\theta_t)) = \frac{\partial \mathcal{L}(\theta)}{\partial d(\theta, x)} \frac{\partial d(\theta, x)}{\partial \theta}$$

Poincaré Embeddings

- La retracción es $\mathcal{R}_\theta(v) = \theta + v$ (natural gradient method)
- Finalmente proyecta sobre $\mathcal{B}^d$:

$$proj(\theta) = \begin{cases} \frac{\theta}{||\theta||} - \epsilon & \text{si } ||\theta|| > 1 \\ \theta & \text{en otro caso} \end{cases}$$

  con ϵ chiquito para asegurar estabilidad numérica.
- Full update:

$$\theta_{t+1} = proj(\theta_t - \eta_t (\frac{1 - ||x||^2}{2})^2 \nabla_E \mathcal{L}(\theta_t))$$

Poincaré Embeddings - Ejemplos

1. Taxonomies: transitive closure of the WORDNET noun hierarchy:
   - Sea $\mathcal{D} = \{(u, v)\}$ conjunto de relaciones *hypernymy* entre sustantivos, se define:

     $$\mathcal{L}(\theta) = \sum_{(\theta_u,\theta_v)\in\mathcal{D}} \log \frac{e^{-d(_u,\theta_v)}}{1 + \sum_{(\theta_u,\theta'_v)\in\mathcal{N}(u)} e^{-d(\theta_u,\theta_{v'})}} \quad \text{con} \quad \mathcal{N}(u) = \{v' : (u, v') \notin \mathcal{D}\}$$

     conjunto de ejemplos negativos
   - Evaluación: Reconstrucción (generando a partir de los embeddings) y Predicción de Enlaces (separando en entrenamiento, validación y test)
   - Métricas: rank y MAP
2. Network embedding: redes de co-autoria
3. Lexical entailment

Poincaré Embeddings - Network Embeddings

1. Redes de co-autoria: assume la misma probabilidad de aristas que antes:

$$P((u,v) = 1|\Theta) = [1 + e^{\frac{\beta}{2}(d(\theta_u,\theta_v)-R)}]^{-1}$$

2. Jerarquia dada por los temas: misma área más probabilidad de tener arista.
3. Los parámetros β y R son hyperparámetros
4. Usa como función de pérdida la entropia cruzada

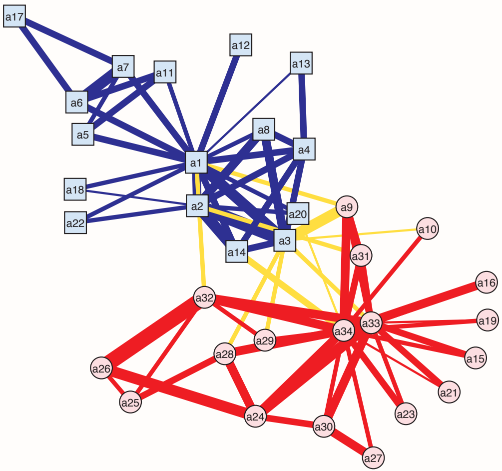
Poincaré Embeddings - Network Embeddings

Table 2: Mean average precision for Reconstruction and Link Prediction on network data.

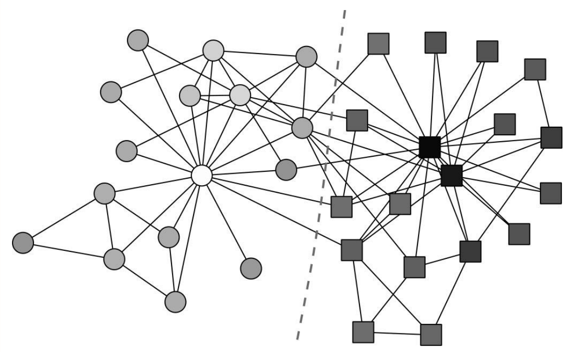
		Dimensionality							
		Reconstruction				Link Prediction			
		10	20	50	100	10	20	50	100
ASTROPH N=18,772; E=198,110	**Euclidean**	0.376	0.788	0.969	0.989	0.508	0.815	0.946	0.960
	Poincaré	0.703	0.897	0.982	0.990	0.671	0.860	0.977	0.988
CONDMAT N=23,133; E=93,497	**Euclidean**	0.356	0.860	0.991	0.998	0.308	0.617	0.725	0.736
	Poincaré	0.799	0.963	0.996	0.998	0.539	0.718	0.756	0.758
GRQC N=5,242; E=14,496	**Euclidean**	0.522	0.931	0.994	0.998	0.438	0.584	0.673	0.683
	Poincaré	0.990	0.999	0.999	0.999	0.660	0.691	0.695	0.697
HEPPH N=12,008; E=118,521	**Euclidean**	0.434	0.742	0.937	0.966	0.642	0.749	0.779	0.783
	Poincaré	0.811	0.960	0.994	0.997	0.683	0.743	0.770	0.774

- Reconstruye mejor que predice (?)
- Comparación justa de dimensiones? impacto de duplicar la dimensión..
- Clustering? otras métricas del grafo mismo?

Ejemplo: Zachary's Karate Club

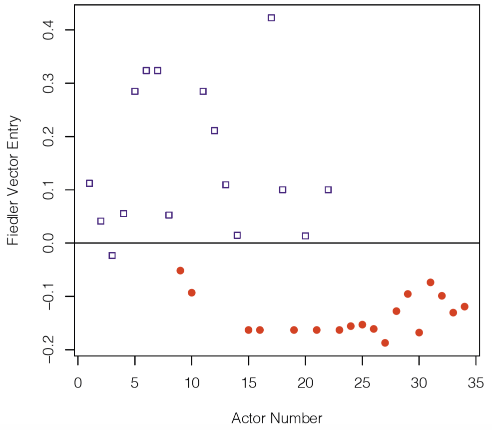


Maximización de la Modularidad



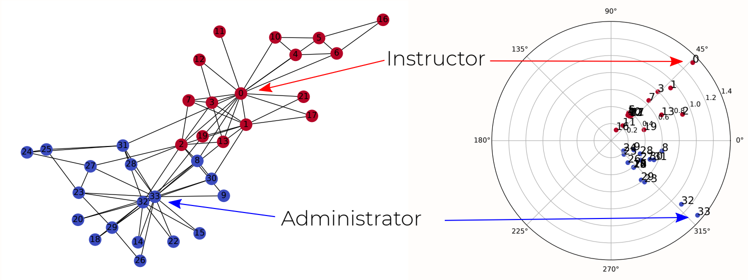
- Maximización espectral de la modularidad
  - forma de los vértices indica pertenencia a la comunidad
  - linea punteada indica la partición encontrada por el algoritmo
  - colores de los vértices indican la fuerza de la pertenencia

Minimización de cortes



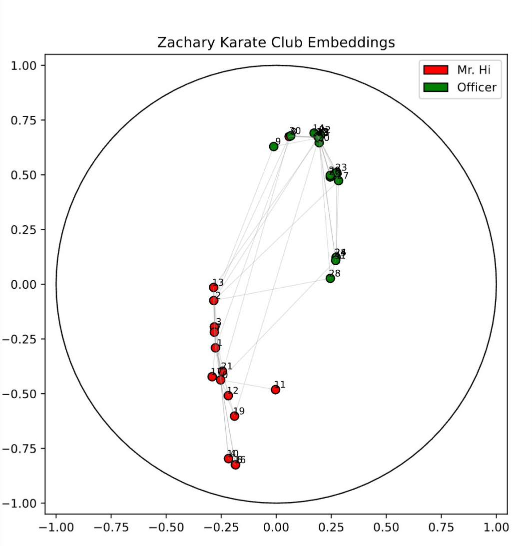
El segundo vector propio más chico v_2 del Laplaciano, verifica que sus entradas suman cero: el signo indica pertenencia a la comunidad.

RDPG Zachary's KC

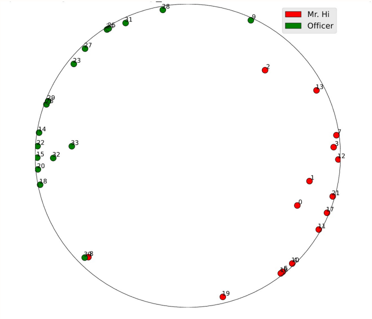


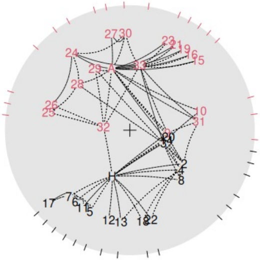
- Método espectral para $d = 2$.
- Fácil de interpretar:
  - $\Rightarrow$ Norma indica qué tan conectado está el nodo
  - $\Rightarrow$ Ángulo indica afinidad

Poincaré Embeddings - Zachary's KC



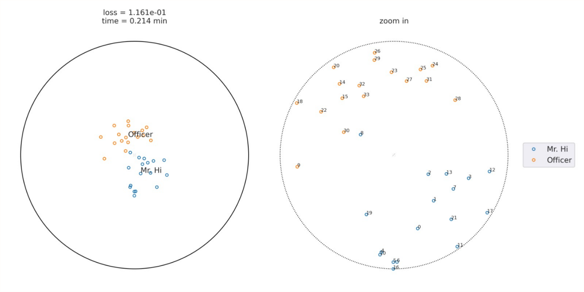
Otros Embeddings Hiperbólicos - Zachary's KC

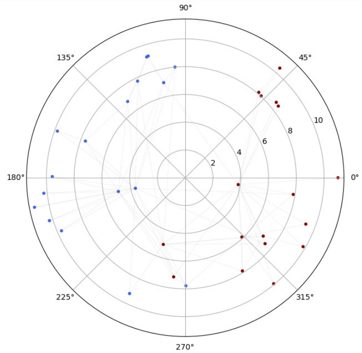




Izquierda: Lorenz Embeddings
Derecha: Hydra

Otros Embeddings Hiperbólicos - Zachary's KC





Izquierda:Poincaré Maps
Derecha: D-Mercator

Random Dot Product Graph - RDPG

Random Dot Product Graphs

- Each node is endowed with a vector $\mathbf{x}$ in a latent space $\mathcal{X}_d \subset \mathbb{R}^d$ such that for all

$$\mathbf{x}, \mathbf{y} \in \mathcal{X}_d \quad \Rightarrow \quad \mathbf{x}^\top \mathbf{y} \in [0,1]$$

- Existence of an edge related to the alignment of the corresponding latent positions:

$$\mathbb{P}(A_{ij} = 1 | \mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$$

Random Dot Product Graphs

- Each node is endowed with a vector $\mathbf{x}$ in a latent space $\mathcal{X}_d \subset \mathbb{R}^d$ such that for all

$$\mathbf{x}, \mathbf{y} \in \mathcal{X}_d \quad \Rightarrow \quad \mathbf{x}^\top \mathbf{y} \in [0, 1]$$

- Existence of an edge related to the alignment of the corresponding latent positions:

$$\mathbb{P}(A_{ij} = 1 | \mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$$

- Stack the latent positions in a matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_v}]^\top \in \mathbb{R}^{N_v \times d}$

$\Rightarrow$ RDPG model specifies that $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$ matrix of probability connections

Young and Scheinerman, "Random dot product graph models for social networks," *WAW*, 2007

Estimation of Latent Positions

- **Q:** Given $G = (V, E)$ from an RDPG, find the 'best' $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_v}]^\top$?

Estimation of Latent Positions

- **Q:** Given $G = (V, E)$ from an RDPG, find the 'best' $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_v}]^\top$?

- **Key:** Observed adjacency matrix $\mathbf{A}$ is a noisy realization of $\mathbf{P}$ ($\mathbb{E}\{\mathbf{A}\} = \mathbf{P}$)

- Suggests a LS regression approach to find $\mathbf{X}$ s.t. $\mathbf{X}\mathbf{X}^\top \approx \mathbf{A}$

$$\hat{\mathbf{X}}_{LS} = \underset{\mathbf{X} \in \mathbb{R}^{N_v \times d}}{\operatorname{argmin}} \|\mathbf{X}\mathbf{X}^\top - \mathbf{A}\|_F^2$$

Estimation of Latent Positions

- **Q:** Given $G = (V, E)$ from an RDPG, find the 'best' $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_v}]^\top$?

- **Key:** Observed adjacency matrix $\mathbf{A}$ is a noisy realization of $\mathbf{P}$ ($\mathbb{E}\{\mathbf{A}\} = \mathbf{P}$)

- Suggests a LS regression approach to find $\mathbf{X}$ s.t. $\mathbf{X}\mathbf{X}^\top \approx \mathbf{A}$

$$\hat{\mathbf{X}}_{LS} = \underset{\mathbf{X} \in \mathbb{R}^{N_v \times d}}{\operatorname{argmin}} \|\mathbf{X}\mathbf{X}^\top - \mathbf{A}\|_F^2$$

- Adjacency spectral embedding (ASE):

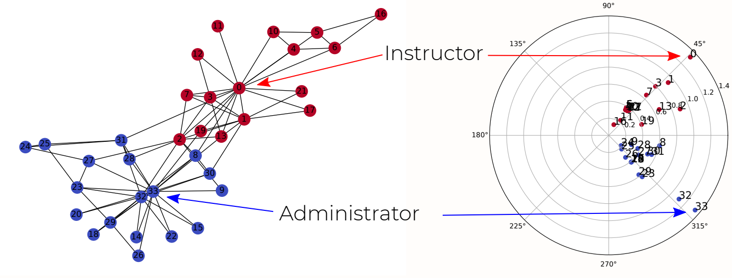
$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \approx \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^\top = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}^{1/2}\hat{\mathbf{\Lambda}}^{1/2}\hat{\mathbf{U}}^\top = \hat{\mathbf{X}}_{LS}\hat{\mathbf{X}}_{LS}^\top$$

  - $\hat{\mathbf{\Lambda}} = \operatorname{diag}(\lambda_1^+, \ldots, \lambda_d^+)$ and $\hat{\mathbf{U}} = [\mathbf{U}_1, \ldots, \mathbf{U}_d]$ ($\lambda^+ := \max(0, \lambda)$)

A. Athreya et al, "Statistical inference on random dot product graphs: A survey," *JMLR*, 2018

Interpretability of the Embeddings

- Example: Zachary's karate club graph with $N_v = 34$, $N_e = 78$ (left)



- Node embeddings (rows of $\hat{\mathbf{X}}_{LS}$) for $d = 2$ (right)
- Interpretability of embeddings a valuable asset for RDPGs
  - $\Rightarrow$ Vector magnitudes indicate how well connected nodes are
  - $\Rightarrow$ Vector angles indicate nodes' affinity