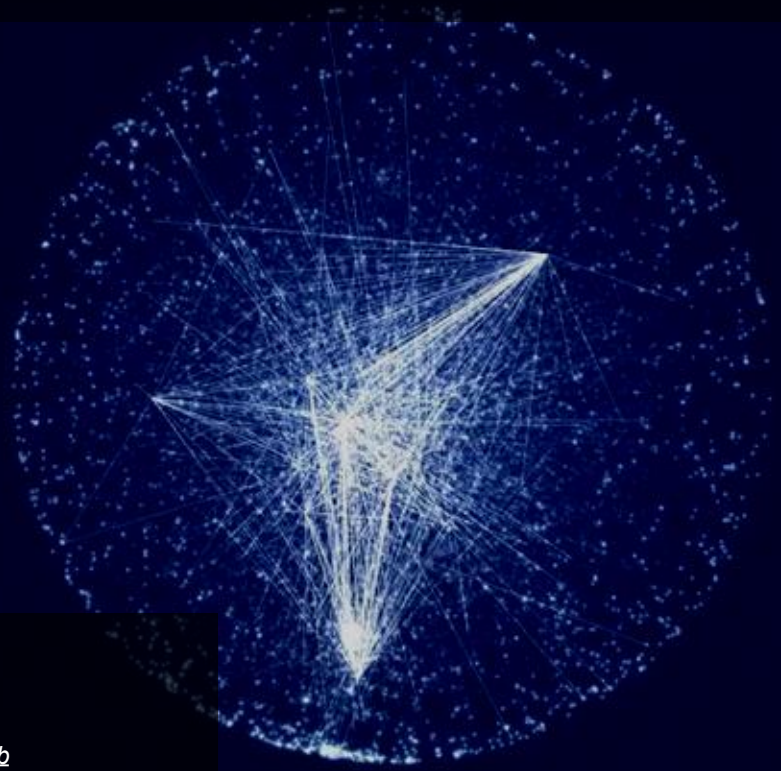
Redes y Sistemas Complejos



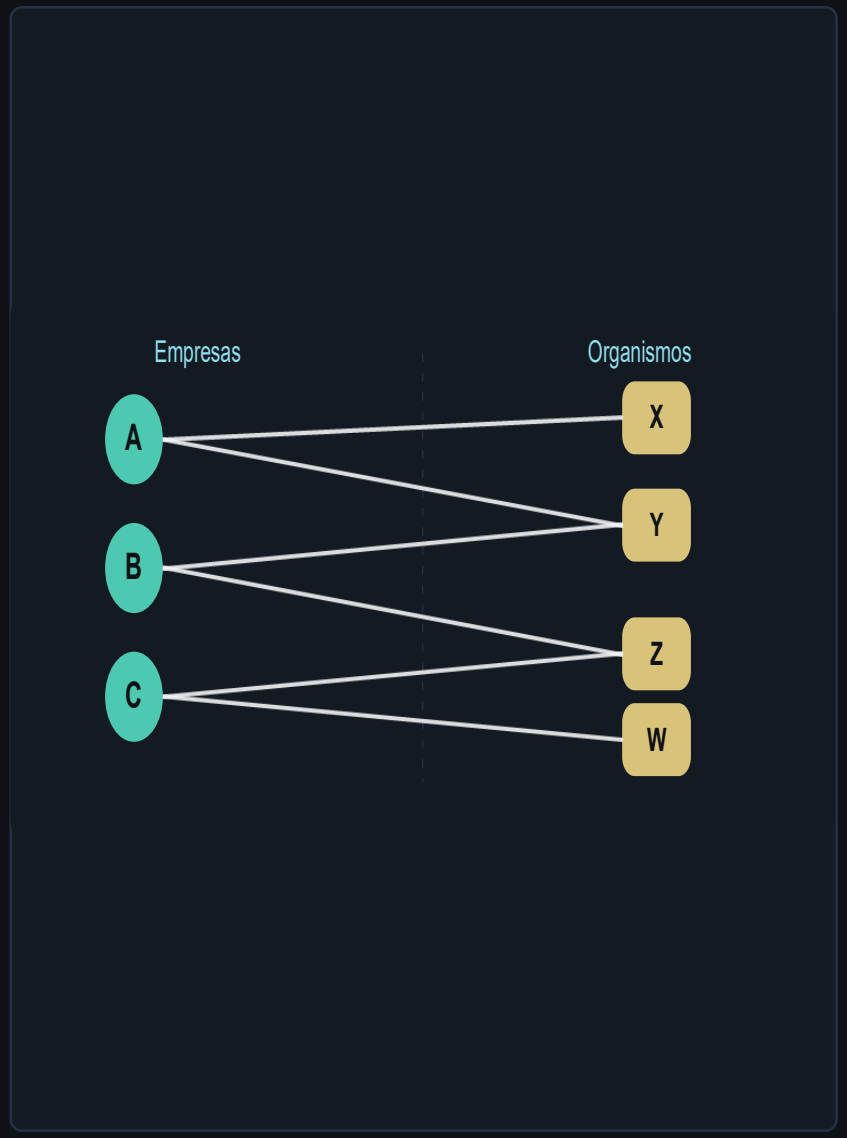
Cristian Candia, Ph.D.

Universidad del Desarrollo (UDD), Chile
Director at the Computational Research in Social Sciences Lab
Associate Professor, Data Science Institute, School of Engineering

Northwestern University, United States.
External Faculty, Northwestern Institute on Complex Systems (NICO)
Kellogg School of Management.

Capybara Spa (AI & Network Science for Preventive, Traceable School Coexistence Compliance)
Founder & Chief Scientific and Technological Officer (CSTO)

Video: Protein-Protein Interactions
Credit: Mauro Martino



Curso Redes y Sistemas Complejos

Clase 2

Redes bipartitas y sesgo por proyección

Representación correcta del sistema, costos de colapsar la estructura e interpretación cauta en Data Science.

Cristian Candia · Instituto de Data Science · UDD

De dónde venimos y hacia dónde vamos

Semana 1

- redes como decisión de representación
- IDs, granularidad y edges mal definidos
- duplicados, self-loops, dirección y peso
- métricas básicas en redes simples

Semana 2

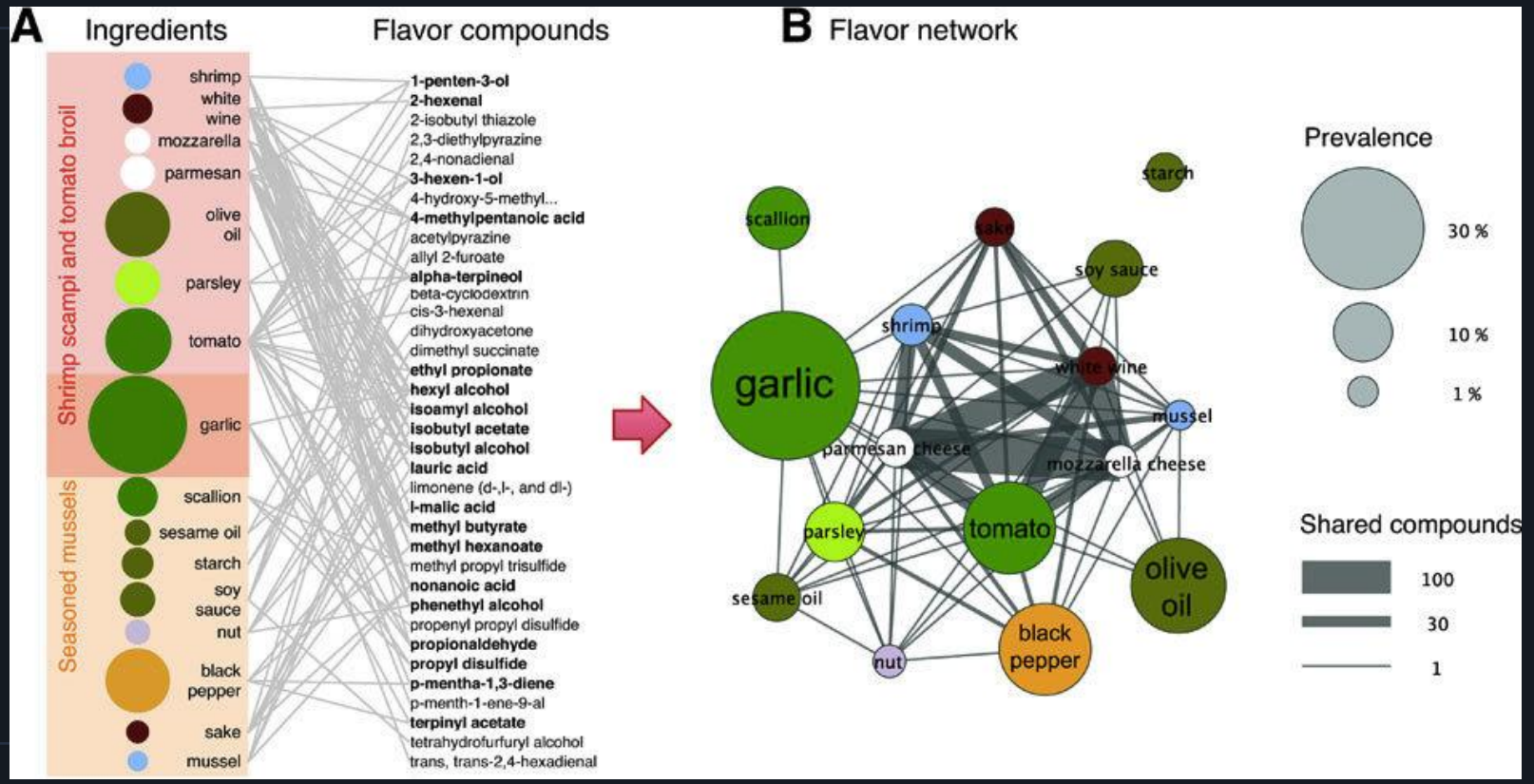
- cuando el sistema correcto es 2-mode
- bipartita como objeto primario
- qué se gana y qué se pierde al proyectar
- cómo evitar inferencias espurias

Idea de continuidad: la representación no sólo organiza los datos; también condiciona la inferencia.

No todo sistema relacional es 1-mode

Ejemplos naturales de redes bipartitas:

- autores ↔ artículos
- actores ↔ películas
- estudiantes ↔ cursos
- usuarios ↔ productos
- enfermedades ↔ genes



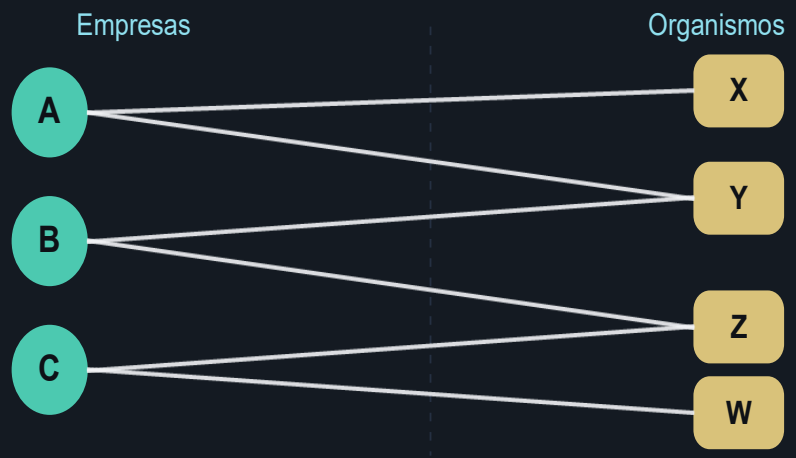
¿Qué es una red bipartita?

Un grafo bipartito es un grafo

$$G = (U, V, E)$$

tal que:

- U y V son conjuntos disjuntos;
- $E \subseteq U \times V$;
- no existen enlaces $U - U$ ni $V - V$.



- dos conjuntos disjuntos de nodos
- las aristas sólo conectan conjuntos distintos
- ejemplo: empresas ↔ organismos
- la relación empresa-empresa no existe en la red original

No siempre debemos “forzar” una red 1-mode.

La representación algebraica natural: matriz de incidencia

Si $|U| = n$ y $|V| = m$, la red bipartita se representa por una matriz rectangular

$$B \in \{0, 1\}^{m \times n}$$

donde

$$B_{ki} = \begin{cases} 1 & si\ v_k \in V\ está\ conectado\ a\ u_i \in U \\ 0 & en\ otro\ caso \end{cases}$$

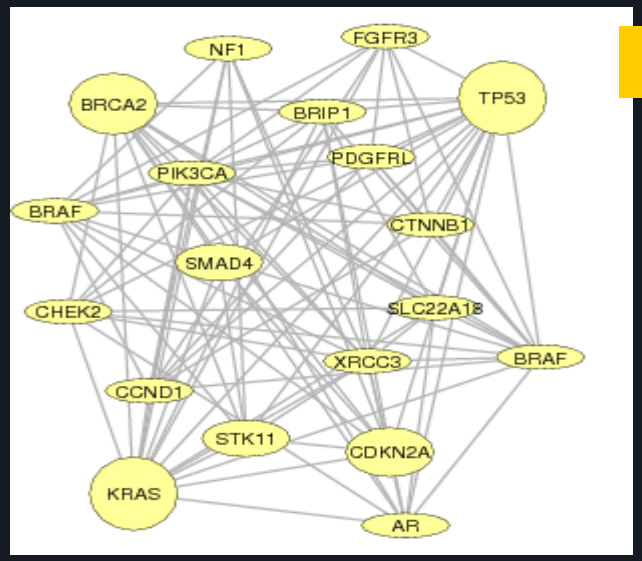
La matriz de incidencia B

Una forma compacta de representar la bipartita antes de proyectar.

	v1	v2	v3	v4
u1	1	1	0	0
u2	0	1	1	0
u3	0	0	1	1

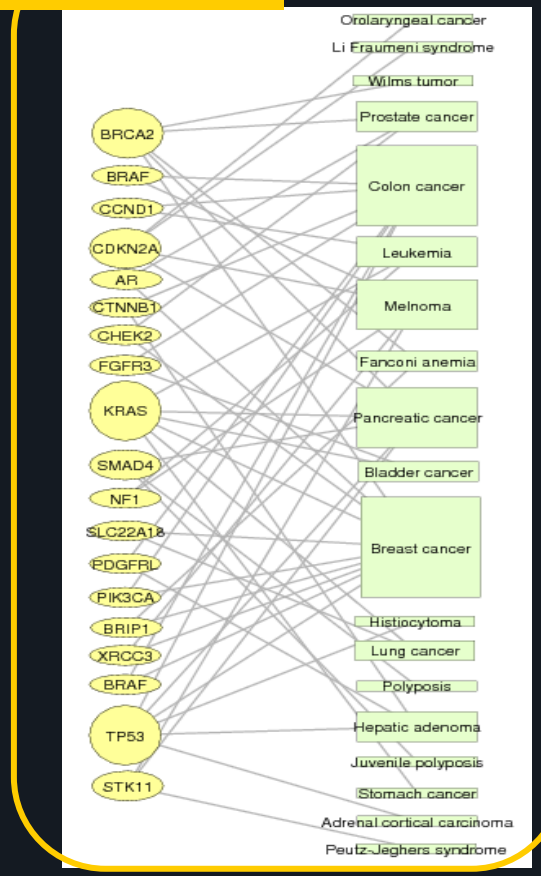
Filas: organismos
Columnas: empresas

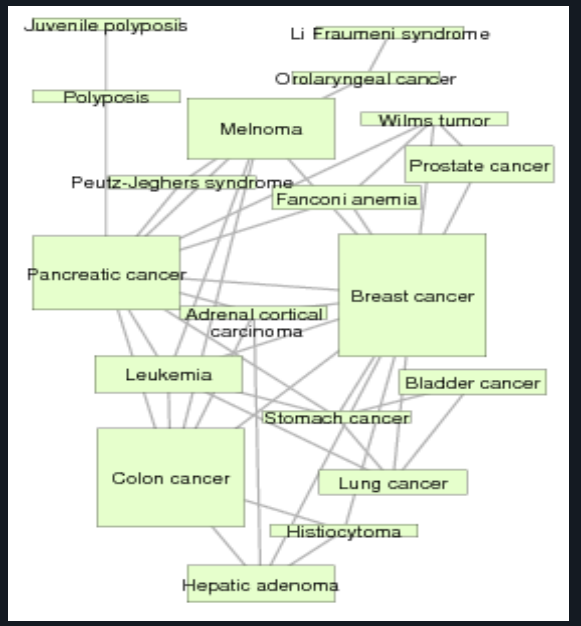
- B codifica presencia/ausencia de vínculo U–V
- la proyección ponderada sobre empresas es B^TB
- la proyección sobre organismos es BB^T
- luego se anula la diagonal



Gene network

DISEASOME PHENOME





Disease network

Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

Lo que es particularmente interesante es la interconexión entre estas dos redes, que nos permite ver cómo alteraciones en genes particulares pueden influir en la susceptibilidad a múltiples enfermedades, lo que demuestra la complejidad de las bases genéticas de las enfermedades humanas.

Grados en una red bipartita

Grado de $u_i \in U$:

$$k_i^{(U)} = \sum_{k=1}^{m} B_{ki}$$

Grado de $v_k \in V$:

$$k_k^{(V)} = \sum_{i=1}^{n} B_{ki}$$

	v1	v2	v3	v4
u1	1	1	0	0
u2	0	1	1	0
u3	0	0	1	1

¿Por qué proyectar?

A veces la pregunta sustantiva recae sobre un solo tipo de entidad:

- colaboración entre autores;
- similitud entre artículos;
- proximidad entre productos;
- solapamiento entre organizaciones.

Entonces construimos una **proyección unipartita**.

Proyección sobre U

Dos nodos $u_i, u_j \in U$ quedan conectados si comparten al menos un vecino en V.

La proyección ponderada se obtiene con:

$$P_U = B^\top B$$

y

$$(P_U)_{ij} = \sum_{k=1}^{m} B_{ki}\, B_{kj}, i \neq j$$

Proyección sobre V

De forma análoga:

$$P_V = BB^{\mathsf{T}}$$

y

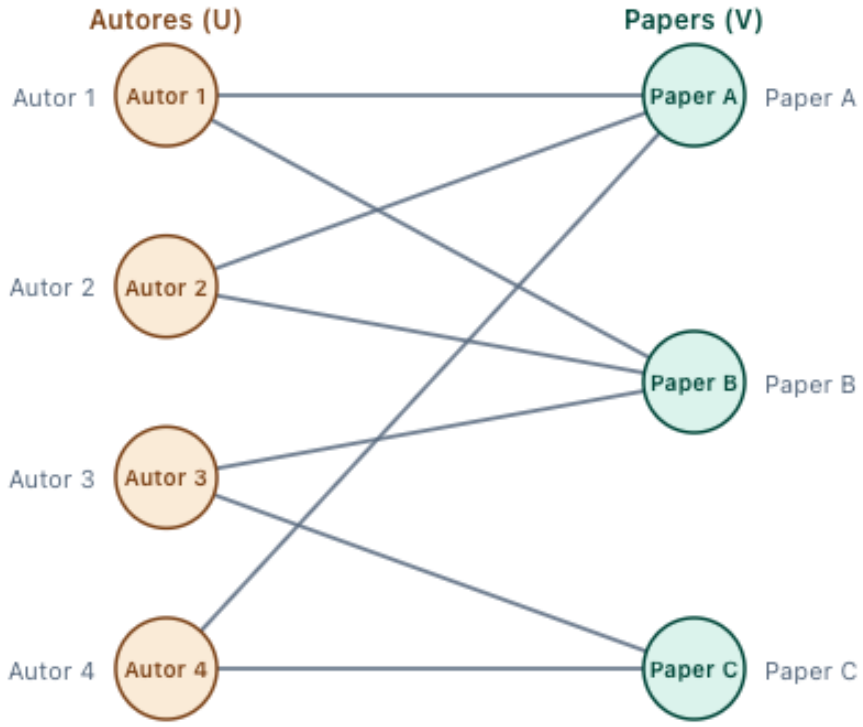
$$(P_V)_{k\ell} = \sum_{i=1}^{n} B_{ki}\, B_{\ell i}, k \neq \ell$$

Qué significa una arista en la proyección

Una arista proyectada entre u_i y u_j no implica interacción directa.

Implica:

- compartir al menos un vecino del otro modo;
- o, en versión ponderada, compartir varios.



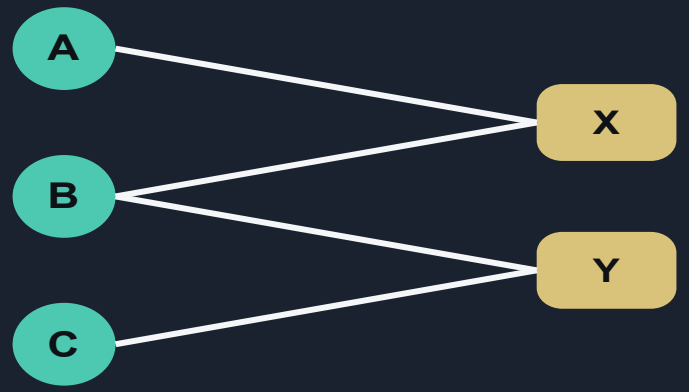
Biadjacencia B

	Paper A	Paper B	Paper C
Autor 1	1	1	0
Autor 2	1	1	0
Autor 3	0	1	1
Autor 4	1	0	1

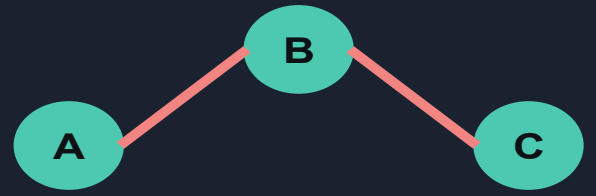
Nodos U	4
Nodos V	3
Aristas |E|	8
Densidad ρ	0.667
$\bar{k}_U$	2.00
$\bar{k}_V$	2.67

Proyectar sirve, pero no es inocente

Bipartita original



Proyección empresa–empresa



Proyección empresa–empresa: dos empresas quedan conectadas si comparten al menos un organismo.
Proyección organismo–organismo: dos organismos quedan conectados si comparten al menos un proveedor.

Qué se pierde al proyectar

Colapsar la estructura hace más simple el análisis, pero más débil la inferencia.

Identidad del mediador

- qué organismo generó la coincidencia
- si el mediador es masivo o específico

Intensidad

- cuántas coincidencias hubo
- si la relación ocurrió una vez o muchas

Estructura completa

- composición de grupos
- tamaño exacto de cada conjunto

Weighted projection ayuda, pero no recupera toda la información.

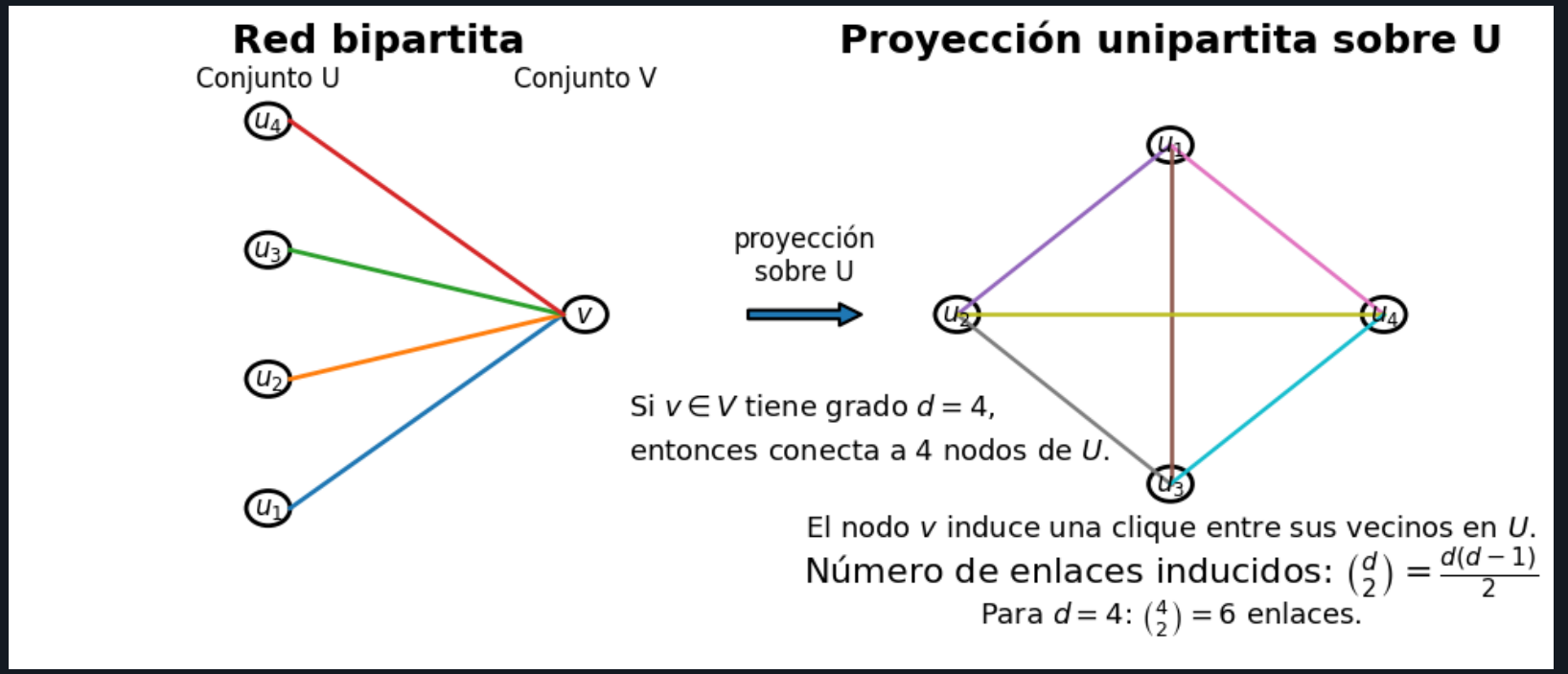
Cada nodo del otro lado induce una clique

Si un nodo de V tiene grado d, entonces induce en la proyección sobre U:

$$\binom{d}{2} = \frac{d(d-1)}{2}$$

enlaces.

- un solo nodo masivo puede densificar toda la proyección
- la centralidad puede reflejar exposición a hubs, no un rol sustantivo
- hay que contrastar siempre contra la bipartita original



Tres problemas clásicos de la proyección

01

Pérdida de información

la proyección borra qué nodo intermedió la relación y colapsa estructura relevante

02

Inflación de enlaces

un nodo muy activo puede inducir una enorme cantidad de enlaces en la proyección

03

Clustering artificial

triángulos y comunidades aparentes pueden surgir por construcción

Conclusión: una proyección mezcla estructura sustantiva y artefactos de representación.

Entonces, ¿hay que evitar siempre la proyección?

No.

La proyección puede ser útil si:
•la pregunta sustantiva está bien definida sobre un solo modo;
•la interpretación es clara;
•aceptamos el costo informacional;
•y contrastamos siempre con la bipartita original.

Antes de analizar, hay que filtrar la proyección

Problema en redes bipartitas grandes:

la proyección binaria pierde información

la proyección puede inflar masivamente el número de enlaces

nodos de alto grado del otro modo inducen cliques

una red proyectada sin filtrado puede volverse poco interpretable.

Paso 1: partir de la bipartita, no de la proyección

Antes de proyectar, mirar al menos:

- tamaño de ambos modos $(|U|, |V|)$
- número total de enlaces $|E|$
- grados en ambos lados
- densidad bipartita
- heterogeneidad y nodos hiperactivos

Paso 2: proyectar ponderado y normalizar

Proyección ponderada sobre U:

$$P_U = B^\top B$$

donde $(P_U)_{ij}$ cuenta vecinos compartidos.

$$(P_U)_{ij} = | N(i) \cap N(j) |$$

N (i) son sol vecinoes de i

cuenta cuántos vecinos del otro modo comparten i y j.

Luego normalizar el solapamiento, por ejemplo con:

$$J(i, j) = \frac{| N(i) \cap N(j) |}{| N(i) \cup N(j) |}$$

Entonces Jaccard responde:
¿Qué fracción del universo total de vecinos de estos dos nodos corresponde a vecinos compartidos?

Idea: no basta con contar coincidencias; hay que corregir por tamaño de vecindarios.

Paso 3: filtrar y conservar sólo enlaces informativos

Para cada par (i, j) ,conservar el enlace sólo si:

$$| N(i) \cap N(j) | \geq k_{\min}$$

y

$$J(i, j) = \frac{| N(i) \cap N(j) |}{| N(i) \cup N(j) |} \geq \tau$$

donde:

- $k_{\min}$:mínimo de vecinos compartidos
- τ: mínimo de similitud relativa

Idea: exigir simultáneamente evidencia absoluta y solapamiento relativo.

Paso 3: filtrar y conservar sólo enlaces informativos

Una regla práctica de filtrado puede exigir:

- mínimo de vecinos compartidos
- similitud mínima (por ejemplo, Jaccard)
- exclusión o tratamiento especial de nodos extremadamente masivos
- análisis posterior sobre la red filtrada, no sobre la proyección bruta

Qué problema evita esta normalización

Evita, al menos parcialmente, que la proyección quede dominada por:

- nodos muy activos,
- grupos muy grandes,
- coincidencias triviales por volumen.

No elimina todos los sesgos, pero sí mejora mucho la interpretabilidad respecto a usar sólo $B^{\top}B$.

Principio: proyectar, ponderar, normalizar y recién entonces analizar.

Para llevarse hoy

1. La bipartita no es sólo un paso intermedio: muchas veces es el objeto correcto.
2. La proyección puede inflar enlaces, clustering y centralidad.
3. La heterogeneidad de actividad es una fuente central de sesgo.
4. **Primero representar bien. Después medir. Finalmente interpretar con cautela.**