Hypergraph Basics: New Tools for Analyzing Text Data in Sociology

Moses Boudourides

Faculty, Graduate Program on Data Science
Northwestern University

Lecture at the Doctoral School of Sociology
University of Bucharest, Bucharest, Romania
May 5, 2025

Northwestern | SCHOOL OF PROFESSIONAL STUDIES

Why analyze text in sociology? (I)

- Texts as Social Artifacts
  - Texts capture how individuals and groups produce and **share meaning.**
  - They reflect **social norms, values, ideologies, and power relations.**
- Theoretical Foundations
  - Central to hermeneutic sociology (e.g., **symbolic interactionism, ethnomethodology**).
  - Key to critical traditions (e.g., **discourse theory, feminist theory, post-structuralism**).
  - Indispensable in the **constructivist orientation** (language is seen not just as representational but as constitutive of social reality).
- Empirical Evidence
  - Texts include **interviews, field notes, media, documents, social media, etc.**
  - They allow access to both everyday **interactions and institutional narratives.**

Why analyze text in sociology? (II)

- Analytical Flexibility
  - Supports both **qualitative** (e.g., thematic, discourse analysis) and **quantitative** (e.g., **content analysis, topic modeling**) methods.
  - Enables **mixed-methods** approaches and triangulation.
- The Digital and Computational Turn
  - **Digitalization** and **big data** have leveraged the relevance of textual analysis.
  - Computational tools (e.g., **NLP, machine learning**) help scale analysis and uncover hidden patterns.
- Sociological Significance
  - Helps explain identity construction, institutional legitimation, cultural shifts, and power dynamics.
  - Essential for studying public discourse, political rhetoric, social movements, and everyday communication.

Beyond the Traditional Bibliometric Approaches

- Co-Authorship Analysis
  - Patterns of institutional and scholars' collaboration.
- Citation Analysis
  - Citation counts and indices (e.g., Journal Impact Factor, h-index, g-index, e-index, m-index etc.).
  - Co-Citation Networks.
  - Networks of Bibliographic Coupling.
- Keyword and Co-Word Analysis
  - Keyword, concepts etc. co-occurrent networks.
  - Emergence of innovation in Science and Technology.
  - Interdisciplinary Fusion or Colonalization.
- Indicators of Inclusion and Influence
  - Altmetrics.
  - Gender Participation.
  - Open Access Types.
  - Reserach Funding.

Limitations of Traditional Bibliometric Networks

- **Dyadic Focus**: Primarily analyzes pairwise relationships (e.g., co-authorship, citations), limiting complexity.
- **Omission of Multi-Entity Relationships**: Often underrepresents the full scope of collaborations involving multiple authors, concepts, or institutions.
- **Limited to Direct Links**: Struggles with higher-order, semantic, or multiplex structures (e.g., shared themes or multiple types of relationships).
- **Small Contextual Insight**: Difficult to fully incorporate the institutional, geographical, or conceptual background of relationships.
- **Less Suited for Multi-Relational Data**: Traditional methods may fall short in capturing complex, layered networks involving multiple types of connections.

Shortcomings of Graph/Network Analysis

- **Pairwise Bias**: Standard graphs model binary ties; Rather than starting from empirical group-level or higher-order interactions, traditional network models tend to infer them post hoc as emergent communities, often overlooking their explicit structural roles.
- **Flattening of Complexity**: Social relationships or shared concepts are reduced to simple edges, obscuring internal structure or role asymmetries.
- **Ambiguity in Multi-Participation**: When multiple entities interact simultaneously (in heterogeneous multimodal networks), the representation becomes ambiguous or fragmented.
- **Limited Expressiveness for Semantic Relations**: Graphs don't naturally model nuanced textual or conceptual proximities derived from language use.
- Nonetheless, **Knowledge Graphs** address network analysis limits by capturing rich, multi-relational social structures.

Hypergraphs as a Viable Alternative

- **Higher-Order Interactions**: Enable the modeling of higher-order connections, reflecting more nuanced and layered social structures.
- **Modeling Multi-Relational Data**: Hypergraphs represent relationships between more than two entities, capturing complex multi-way interactions (e.g., a document, actors, and concepts in one hyperedge).
- **No Oversimplification**: Hypergraphs avoid flattening social dynamics by keeping the higher-order structure intact, providing more accurate representations of collective behavior.
- **Contextual Representation**: By including multiple participants in a single hyperedge, hypergraphs maintain the contextual richness of sociological phenomena.

Three Examples of Sociologically Relevant Text Data for Hypergraph Analysis

- **Books**, encyclopedias, manifestos, diaries, biographies, speeches, interviews, evaluations etc.:
  - *Hyperedges*: Chapters (or paragraphs or sentences), articles (or items), section, diary entries, events (or life milestones) etc.
  - *Nodes*: Individuals, groups, or institutions involved, concepts, (key)words, places, actions, the embedded background etc.
- **Social media** posts, online forums:
  - *Hyperedges*: Social media posts or messages etc.
  - *Nodes*: All of the above.
- **Films** (or musical compositions, and other forms of artistic expression):
  - *Hyperedges*: Artistic works (e.g., films, theatrical productions, musical compositions),
  - *Nodes*: Directors, composers, performers (cast), genres, stories, narrative elements etc.

So, What Is a Formal Hypergraph?

Hypergraphs are generalizations of graphs, in which edges consist of any number of nodes.

Conventions in interchangeable terminology:

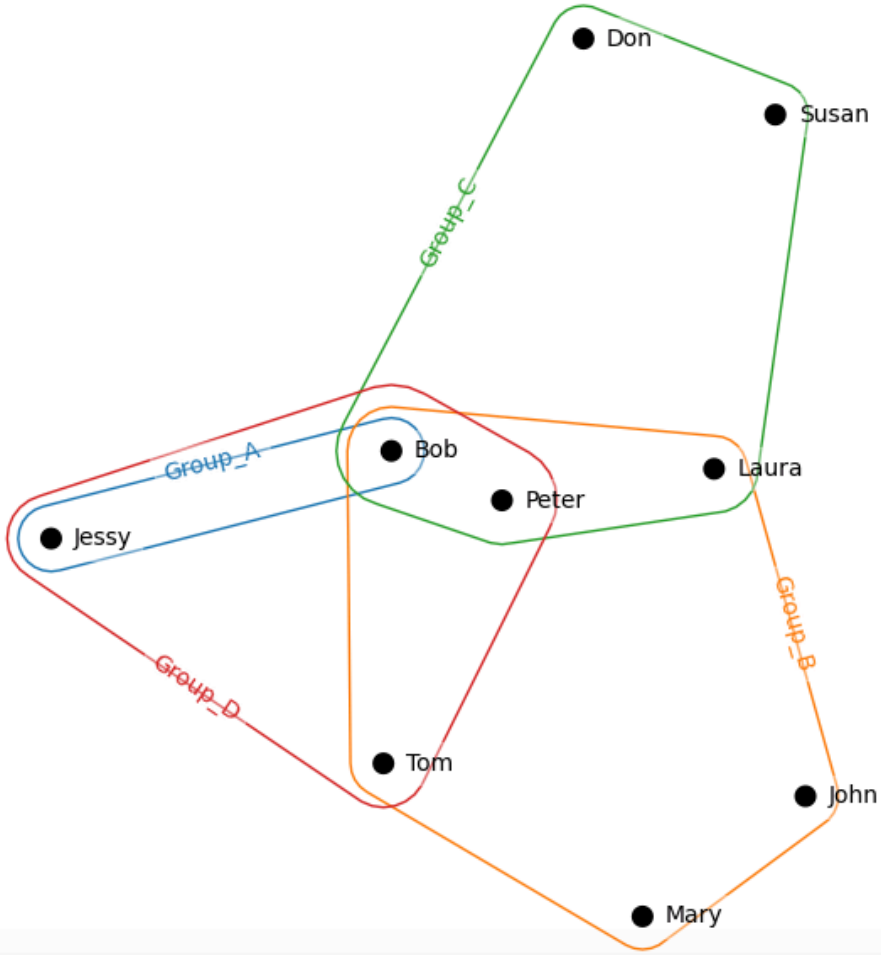
- **graph–network** and **hypergraph–hypernetwork**
- **vertex–node**
- **edge–link** and **hyperedge–hyperlink**
- **bipartite graph** and **two–mode network**

A Mathematical Result: *There is a bijection between hypergraphs and bipartite graphs, as each hypergraph corresponds uniquely to a bipartite graph and vice versa.*

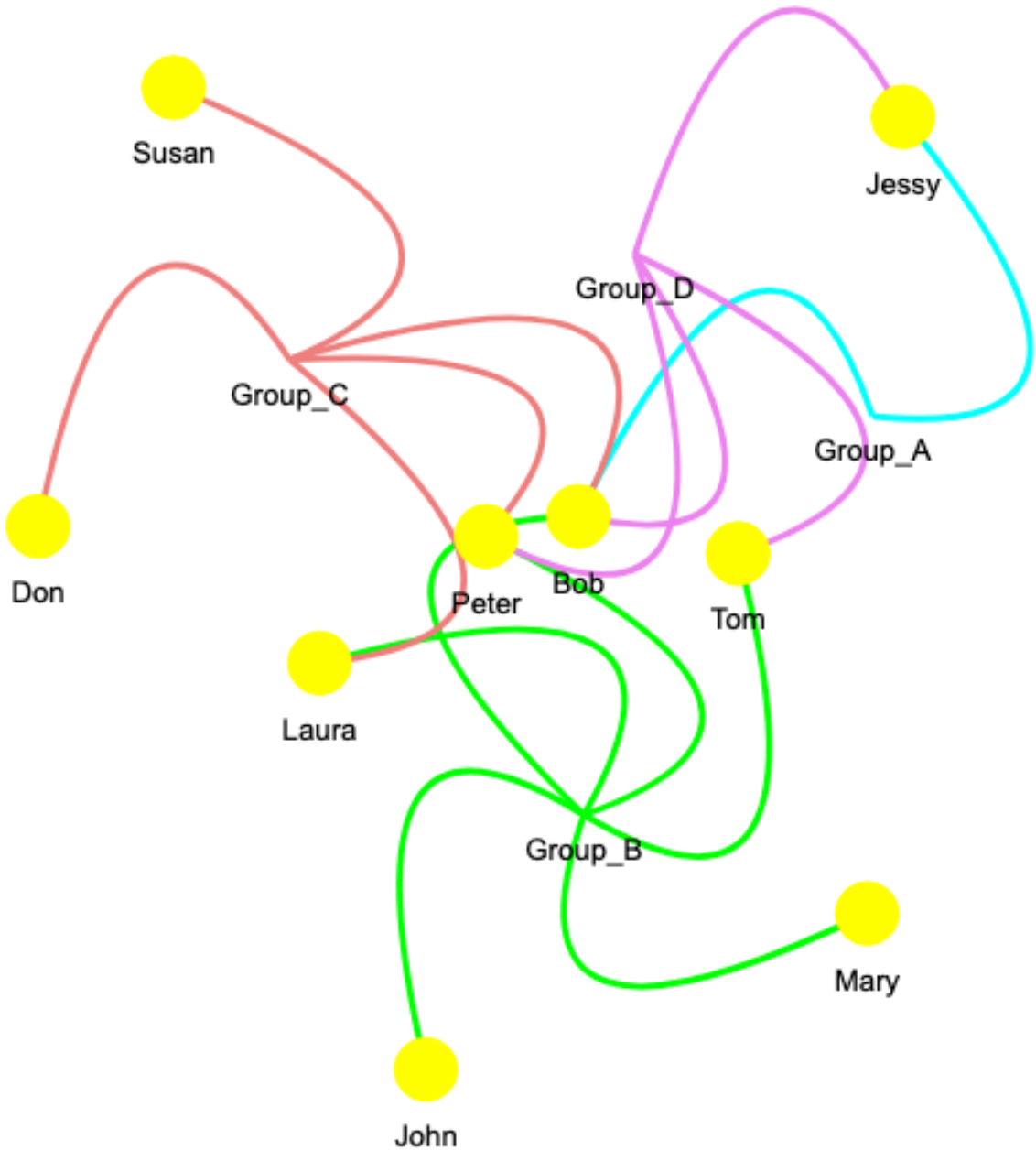
- Although formally equivalent to bipartite graphs,
- Hypergraphs offer a more intuitive and analytically direct representation,
- Avoid reducing multi-entity relations to dyadic structures,
- Preserve the integrity of group-level interactions in sociological data.

Table and Euler Plot of a Hypergraph

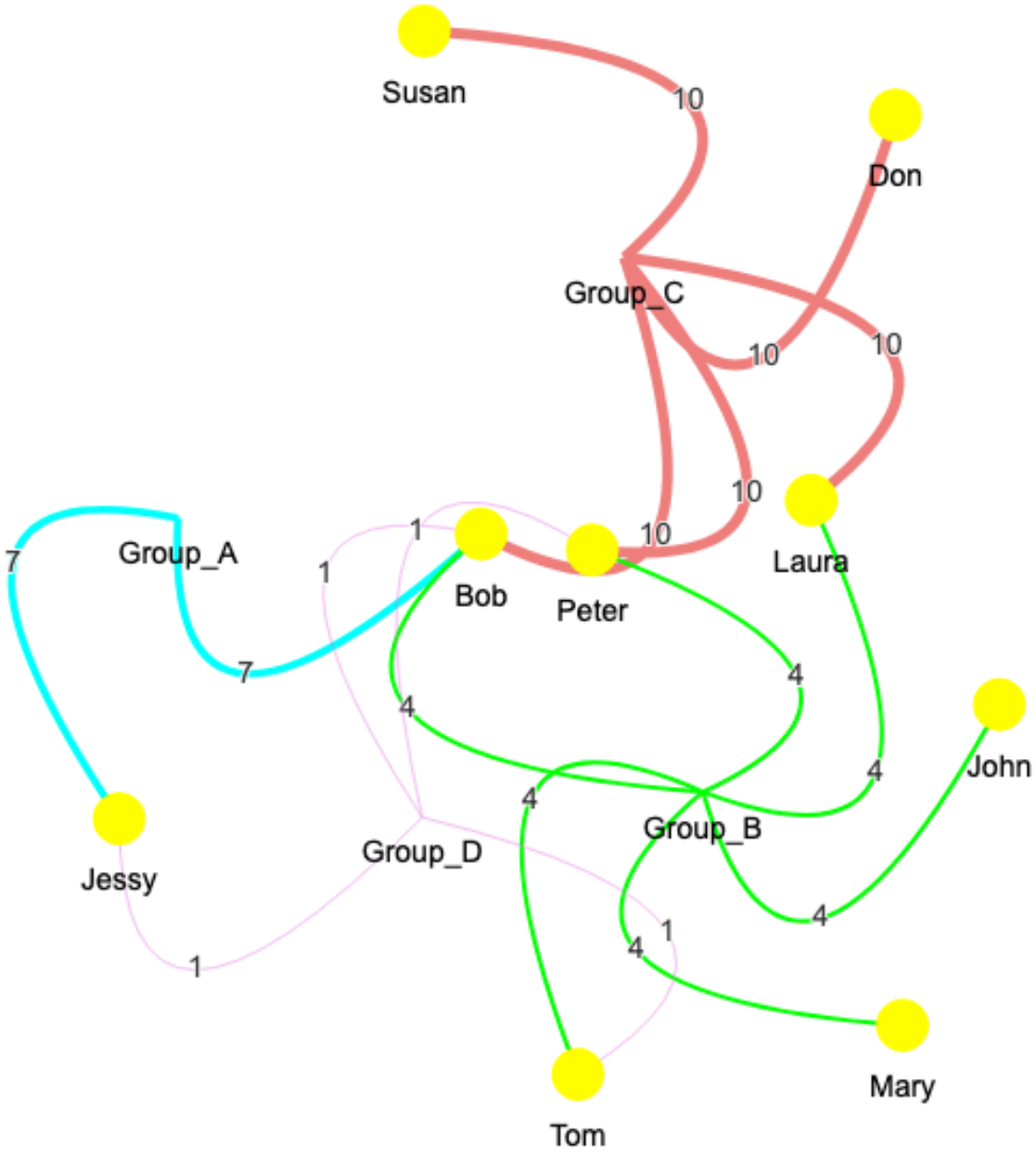
hyperedge	vertices
Group_A	Bob, Jessy
Group_B	Bob, John, Laura, Mary, Peter, Tom
Group_C	Bob, Don, Laura, Peter, Susan
Group_D	Bob, Jessy, Peter, Tom



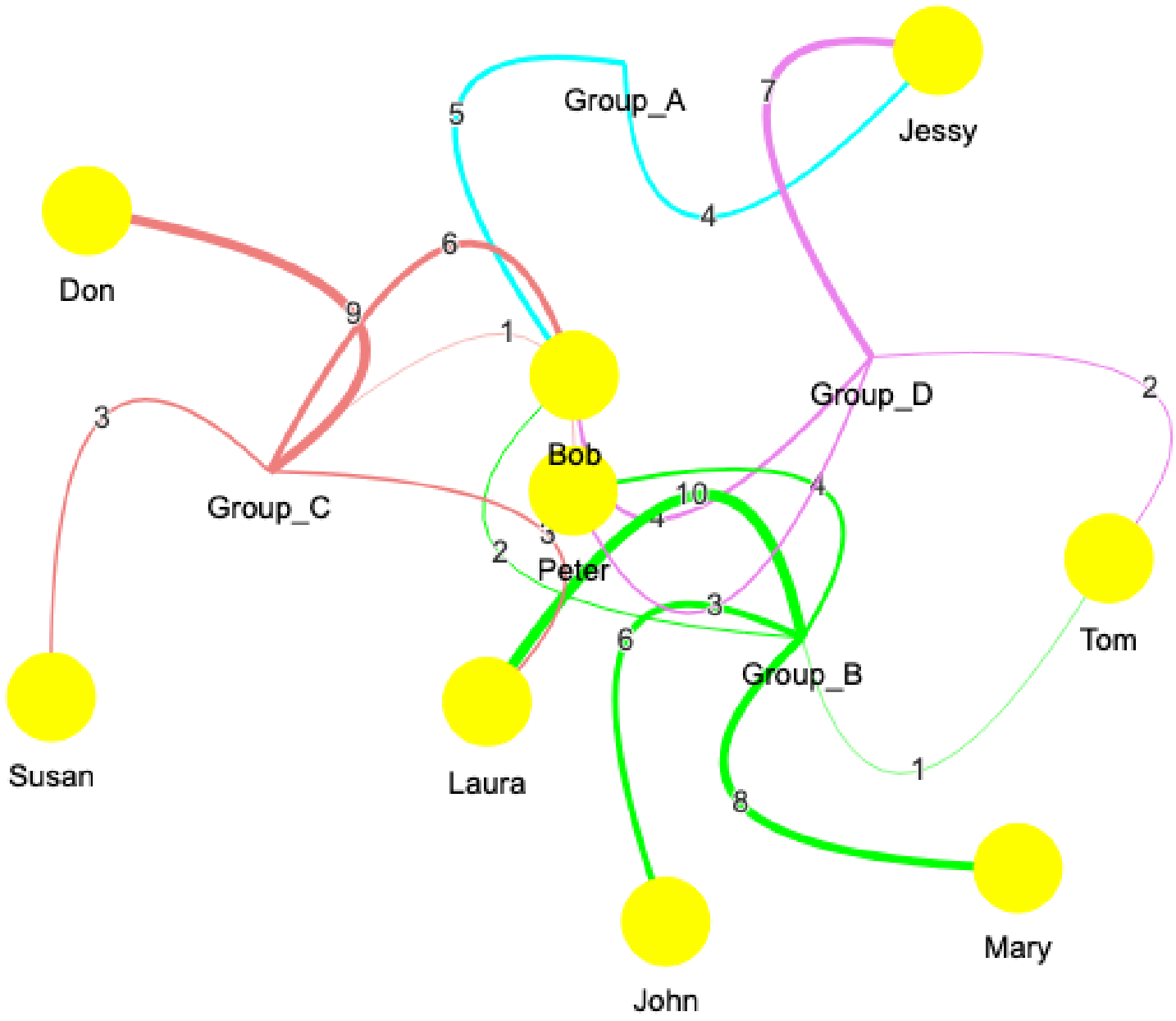
Bipartite Plot of a Hypergraph



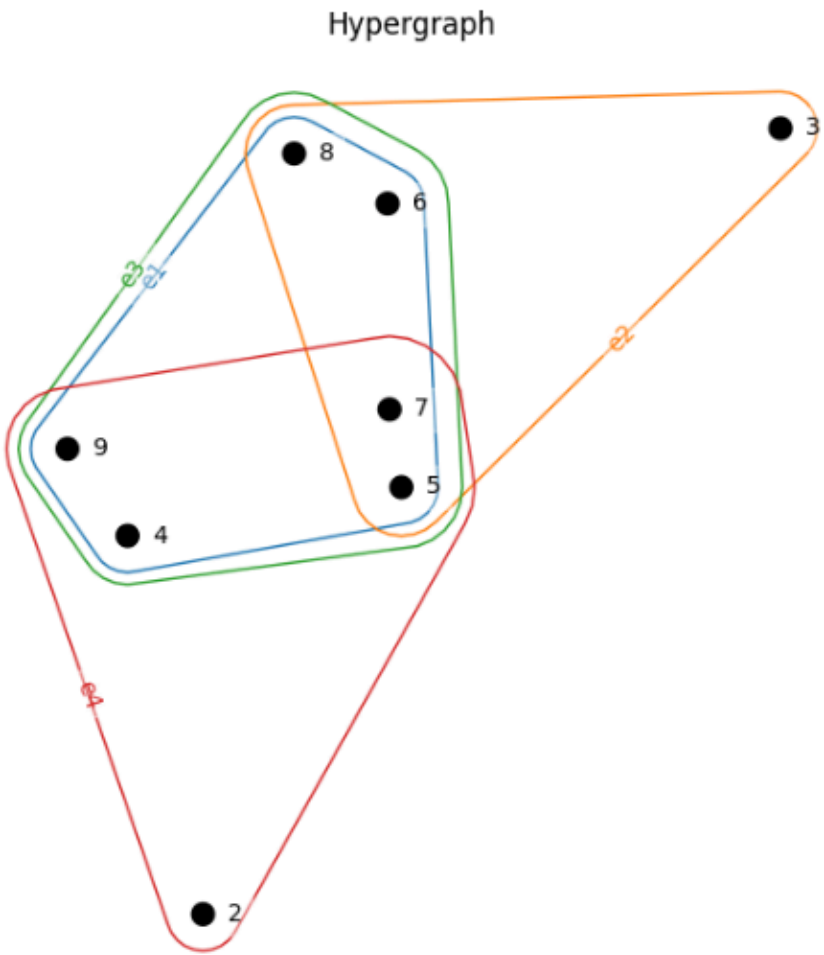
Bipartite Plot of a Multiple Hypergraph

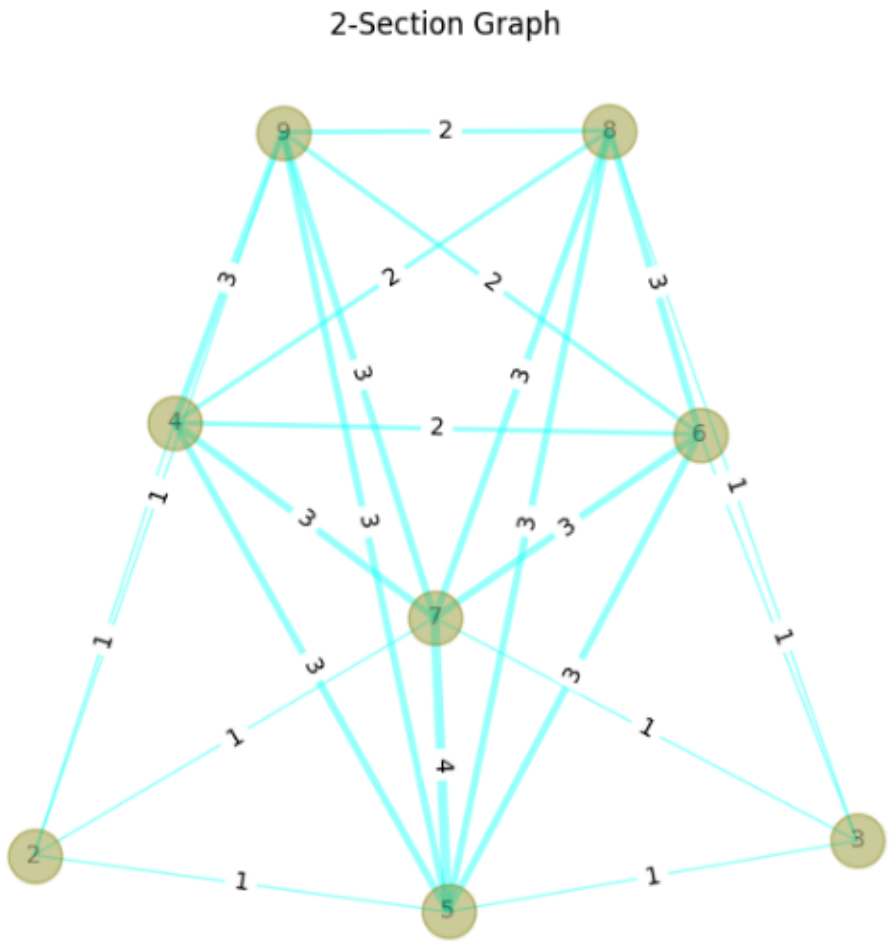


Bipartite Plot of a Hypergraph with Edge-Dependent Vertex Weights (EDVW)



The Clique Expansion of a Hypergraph

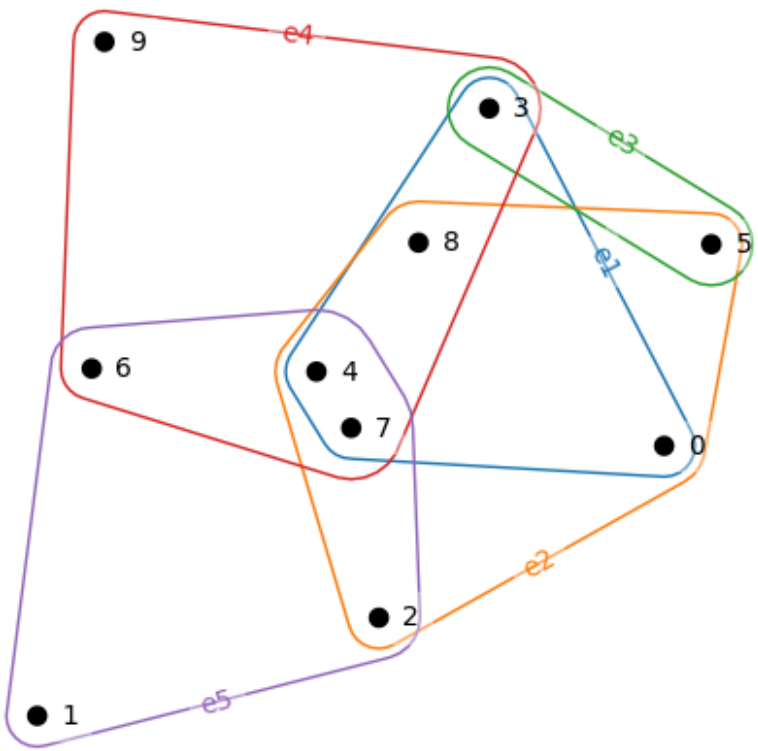




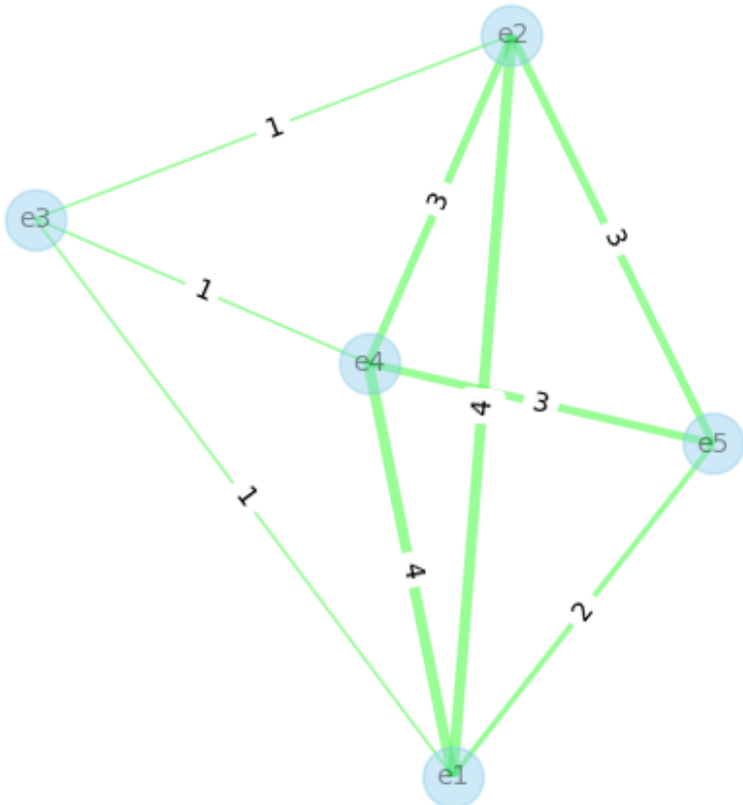
The Line Graph of a Hypergraph

```
Hypergraph Data:
{'e1': [0, 3, 4, 7, 8], 'e2': [0, 2, 4, 5, 7, 8], 'e3': [3, 5], 'e4': [3, 4, 6, 7, 8, 9], 'e5': [1, 2, 4, 6, 7]}
```

Hypergraph



Line Graph

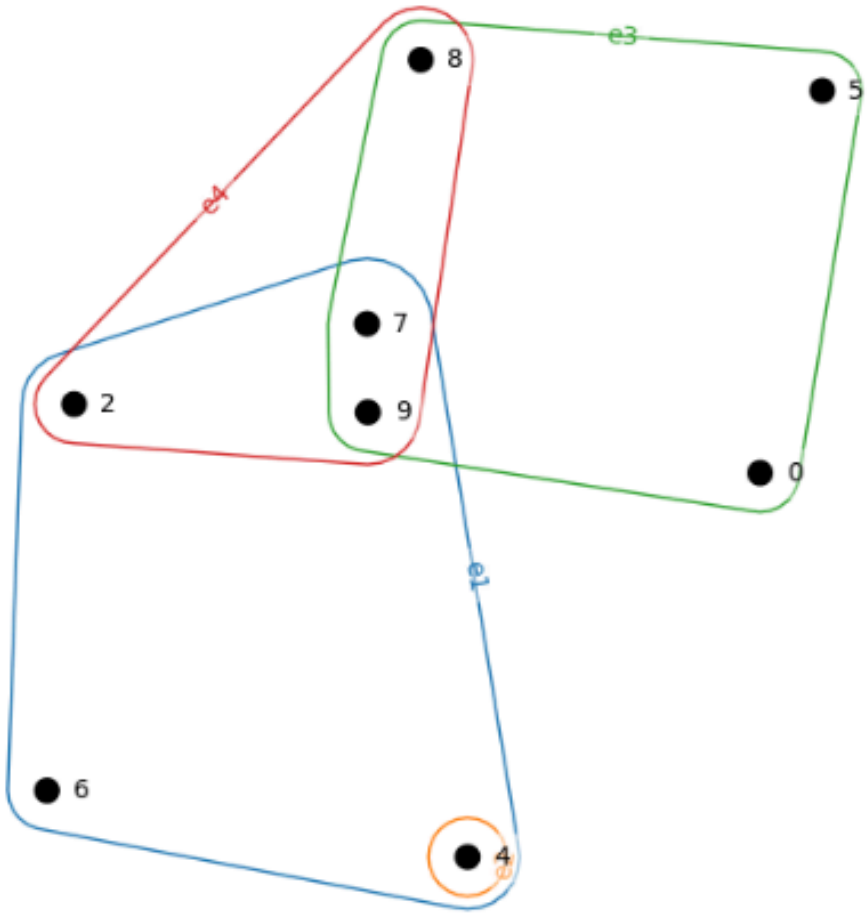


Degrees of Hyperedges and Nodes

- The **degree of hyperedge** e, denoted as $\delta(e)$, is defined as $|e|$, i.e., as the cardinality of (i.e., the number of incident nodes to) e.
- If each hyperedge has the same degree k, the hypergraph is called **uniform** or k-**uniform**.
- The **the degree of node** v, denoted as $\mathbf{deg}(v)$, is defined as the number of hyperedges incident to v:
- If each node has the same degree, the hypergraph is called **regular** or k-**regular**, i.e., if, for every $v \in V, \mathbf{deg}(v) = k$.
- Degree equality:

$$\sum_{v \in V} \mathbf{deg}(v) = \sum_{e \in E} \delta(e).$$

Hypergraph Nodal and Hyperedge Degrees



Hypergraph Data:
{'e1': [2, 4, 6, 7, 9], 'e2': [4], 'e3': [0, 5, 7, 8, 9], 'e4': [2, 7, 8, 9]}

Incidence Matrix:

vertex	e1	e2	e3	e4
0	0	0	1	0
2	1	0	0	1
4	1	1	0	0
5	0	0	1	0
6	1	0	0	0
7	1	0	1	1
8	0	0	1	1
9	1	0	1	1

Node Degrees:
Degree of node 2: 2
Degree of node 4: 2
Degree of node 6: 1
Degree of node 7: 3
Degree of node 9: 3
Degree of node 0: 1
Degree of node 5: 1
Degree of node 8: 2

Hyperedge Degrees:
Degree of hyperedge e1: 5
Degree of hyperedge e2: 1
Degree of hyperedge e3: 5
Degree of hyperedge e4: 4

Hypergraph Connectedness: Traversals

- A **walk** of length ℓ from node u to node v is an alternating node–hyperedge sequence
$$u = v_1, e_1, v_2, e_2, \ldots, v_\ell, e_\ell, v_{\ell+1} = v$$
such that consecutive nodees are distinct ($v_i \neq v_{i+1}$), and each hyperedge e_i is incident to its adjacent nodes v_i and v_{i+1}.
- If the sequence of pairs of consecutive nodes–hyperedges, $(v_1, e_1), (v_2, e_1), (v_2, e_2), \ldots, (v_\ell, e_\ell), (v_{\ell+}1, e_\ell)$, has no repeated hyperedges, then the walk is called a **trail**; note that, in graphs, this corresponds to a walk in which no edge appears more than once, though nodes may be revisited.
- A walk in which all the nodes are distinct is called a **path**. Unlike in simple graphs, where the requirement for distinct nodes ensures that all traversed edges are also distinct, this constraint does not apply to hypergraphs, where paths may include repeated hyperedges.
- Finally, if $u = v_1 = v_{\ell+1} = v$ the path is called a **cycle**.

Hypergraph Connectedness: Distances and Connected Components

- A hypergraph is **connected** if, for every pair of nodes, there exists at least one path connecting them.
- The **distance** between two distinct nodes u and v is the length of the shortest path connecting them, and it is denoted by $d(u, v)$. If no such path exists, their distance is defined as infinite.
- The **eccentricity** of a node is the maximum distance from that node to any other node.
- The **diameter** of the hypergraph is the maximum eccentricity of any node, i.e., it is the maximum distance between any pair of nodes.
- If a hypergraph is not connected, its **connected components** are the equivalence classes induced by the equivalence relation $\mathcal{R}$ on the set of nodes, defined as:

$$(u, v) \in \mathcal{R} \iff \exists \text{ a path between } u \text{ and } v.$$

Hypergraph Centralities

Let $G = (V, E)$ a hypergraph with n nodes and $v \in V$.

- The **closeness centrality** of node v is:

$$\textbf{Closeness}(v) = \frac{n-1}{\sum_{u \neq v} d(u,v)}.$$

- The **harmonic closeness centrality** of node v is:

$$\textbf{Harmonic}(v) = \sum_{u \neq v} \frac{1}{d(u,v)} \cdot \frac{2}{(n-1)(n-2)}.$$

- The **betweenness centrality** of node v is:

$$\textbf{Betweenness}(v) = \frac{1}{\binom{n-1}{2}} \sum_{\substack{s,t \in V \\ s \neq t}} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}},$$

  where $\sigma_{s,t}$ is the total number of shortest paths between nodes s and t, and $\sigma_{s,t}(v)$ is the number of those paths that pass through node v

The Tudisco & Higham Nonlinear Eigenvector Centrality

- Let $G = (V, E)$ be a connected hypergraph with node set $V = \{1, \ldots, n\}$, and let $f : \mathbb{R} \to \mathbb{R}$ be a nonlinear real-valued function. Define a centrality score vector $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, called the **nonlinear singular vector** (**NSV**), as the solution to the nonlinear eigenvalue problem

$$Hf(x) = \lambda x,$$

where the operator $H : \mathbb{R}^n \to \mathbb{R}^n$ is defined component-wise by

$$(Hf(x))_i = \sum_{\substack{e \in E \\ i \in e}} \sum_{j \in e} f(x_j).$$

- If $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is order-preserving and homogeneous of degree less than 1, then there exists a positive eigenvector $x \in \mathbb{R}^n_{>0}$ corresponding to the eigenvalue $\lambda = 1$, i.e., satisfying the fixed-point equation

$$Hf(x) = x.$$

Convergence of the NSV Centrality

- If, in addition, f is strictly concave, then the eigenvector x is unique up to scaling and it can be calculated through the iteration

$$x^{(k+1)} = \frac{Hf(x^{(k)})}{||Hf(x^{(k)})||_2}$$

  starting from a positive initial vector $x^{(0)}$ using the Euclidean L^2–norm $||\cdot||_2$.
- The case $f(x) = x$ reduces to a standard linear eigenvector problem, which generalizes eigenvector centrality from graphs to hypergraphs.
- In the superlinear case, where $f(x) = x^\alpha$ with $\alpha > 1$, centrality scores are biased toward higher-degree nodes, while in the sublinear case, where $f(x) = x^\alpha$ with $0 < \alpha < 1$, centrality is more evenly distributed among nodes.
- Additionally, the logarithmic form $f(x) = \mathbf{log}(x + 1)$ is more appropriate for hypergraphs where a small number of nodes dominate the centrality scores.

Challenges and Future Directions

- Textual data collection tasks
- Computational cost
- Combine with NLP, topic modeling, fieldwork
- Visualization issues
- Interpretation: Sociological meaning of hypergraph analysis
- Egocentered subhypergraph analysis
- Assortativity in hypergraphs
- **Hyperlink prediction** in hypergraphs

Conclusion

- Hypergraphs = powerful new tool
- Well-suited to unlock complex structure in textual data
- Useful for semantic analyses of meaning, discourse, and identity
- **Let's explore this together!!!**

Questions & Discussion

Thank you!

Questions or suggestions?