

CONNECTIONS BETWEEN SOCIAL NETWORK ANALYSIS AND GRAPH ANALYTICS

M.TAMER ÖZSU

UNIVERSITY OF WATERLOO

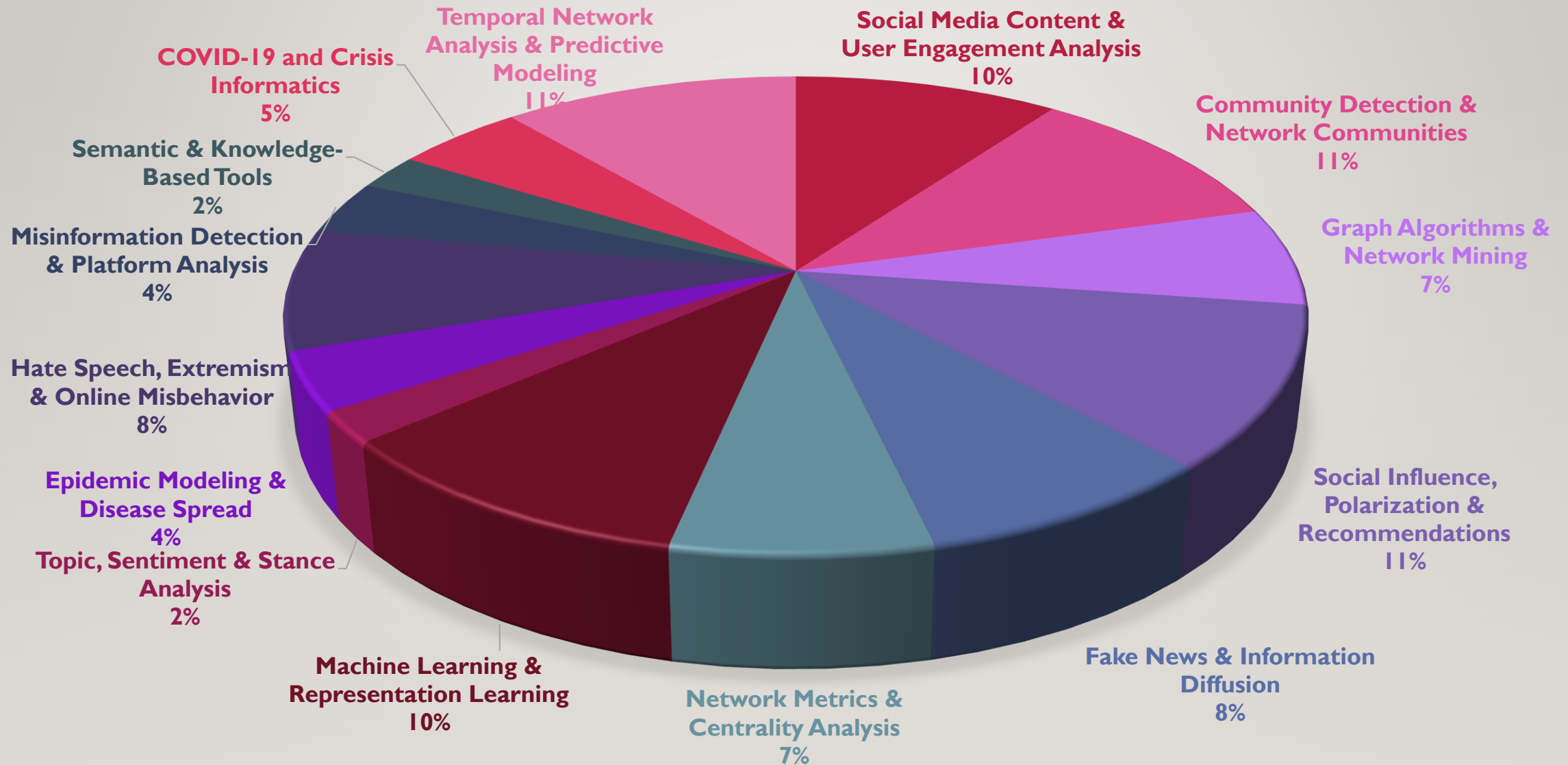
[HTTPS://CS.UWATERLOO.CA/~TOZSU](https://cs.uwaterloo.ca/~tozsu)

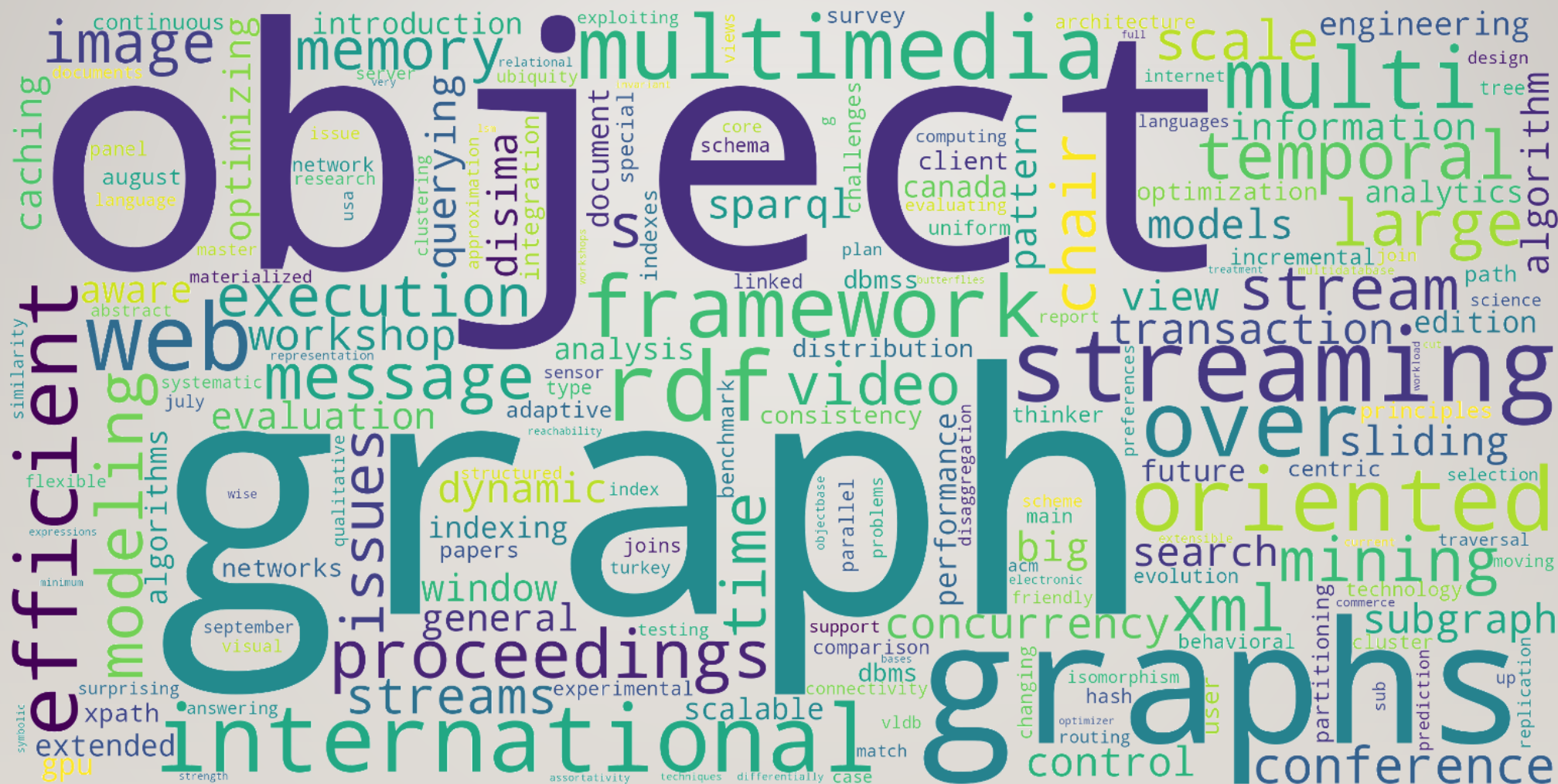


UNIVERSITY OF
WATERLOO



ASONAM INTERESTS (BASED ON 2021-2024 CONFERENCES)



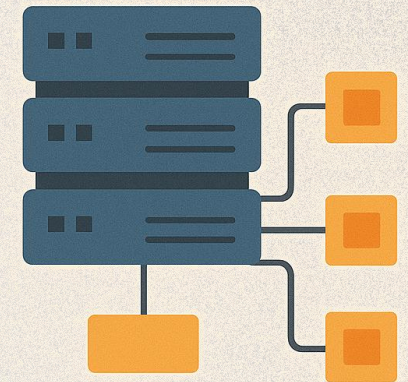


WHY CONNECT SOCIAL NETWORK ANALYSIS & GRAPH ANALYTICS?

- **Social networks are ubiquitous:** From friendships to global supply chains.
- **Social network analysis (SNA):**
Interpreting social structure
 - Focus on meaning, context, and human behaviour.
 - Traditionally, on smaller scales.
- **Graph Analytics (GA):** **Computing at scale**
 - Focus on algorithms, efficiency, and massive datasets.
 - Can be abstract, data-driven.
- **The thesis: Mutual enrichment**
 - Combine SNA's interpretability with the scalability of GA.
 - Uncover patterns in big networks *without losing social meaning*.
 - Use social theories to guide big-data algorithms.



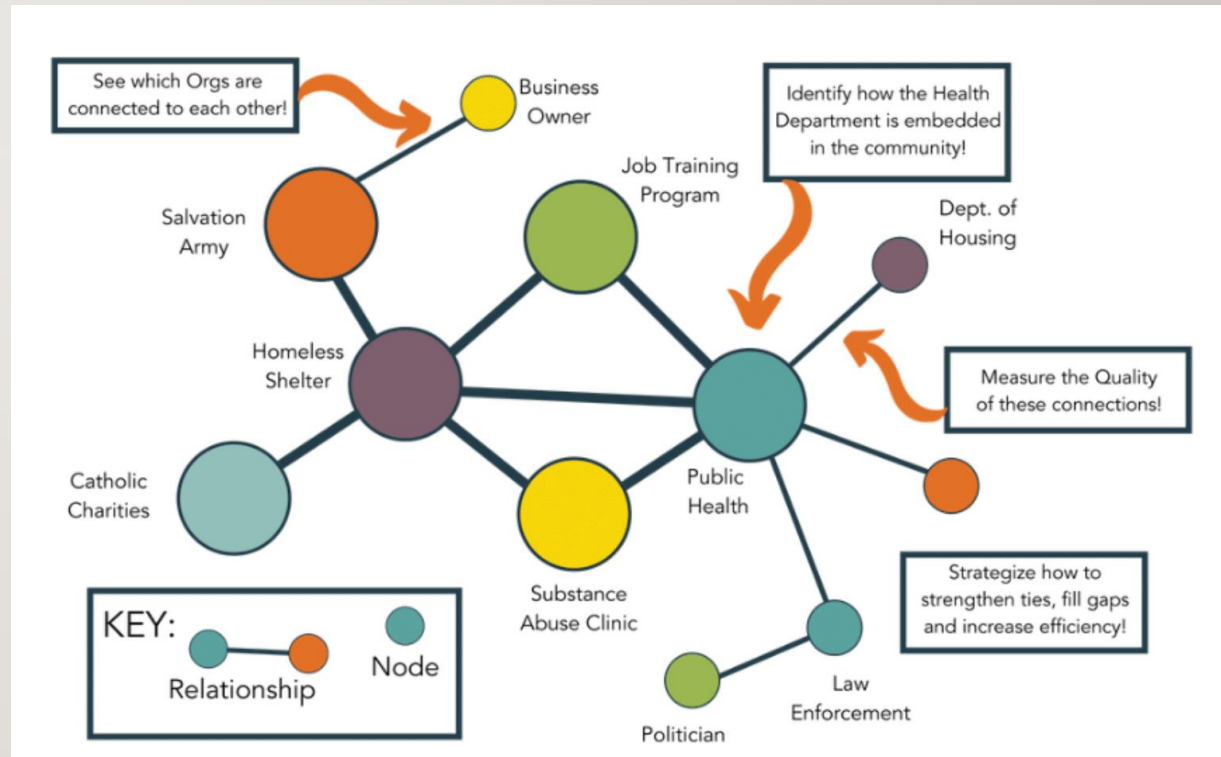
**SOCIAL
NETWORK
ANALYSIS**



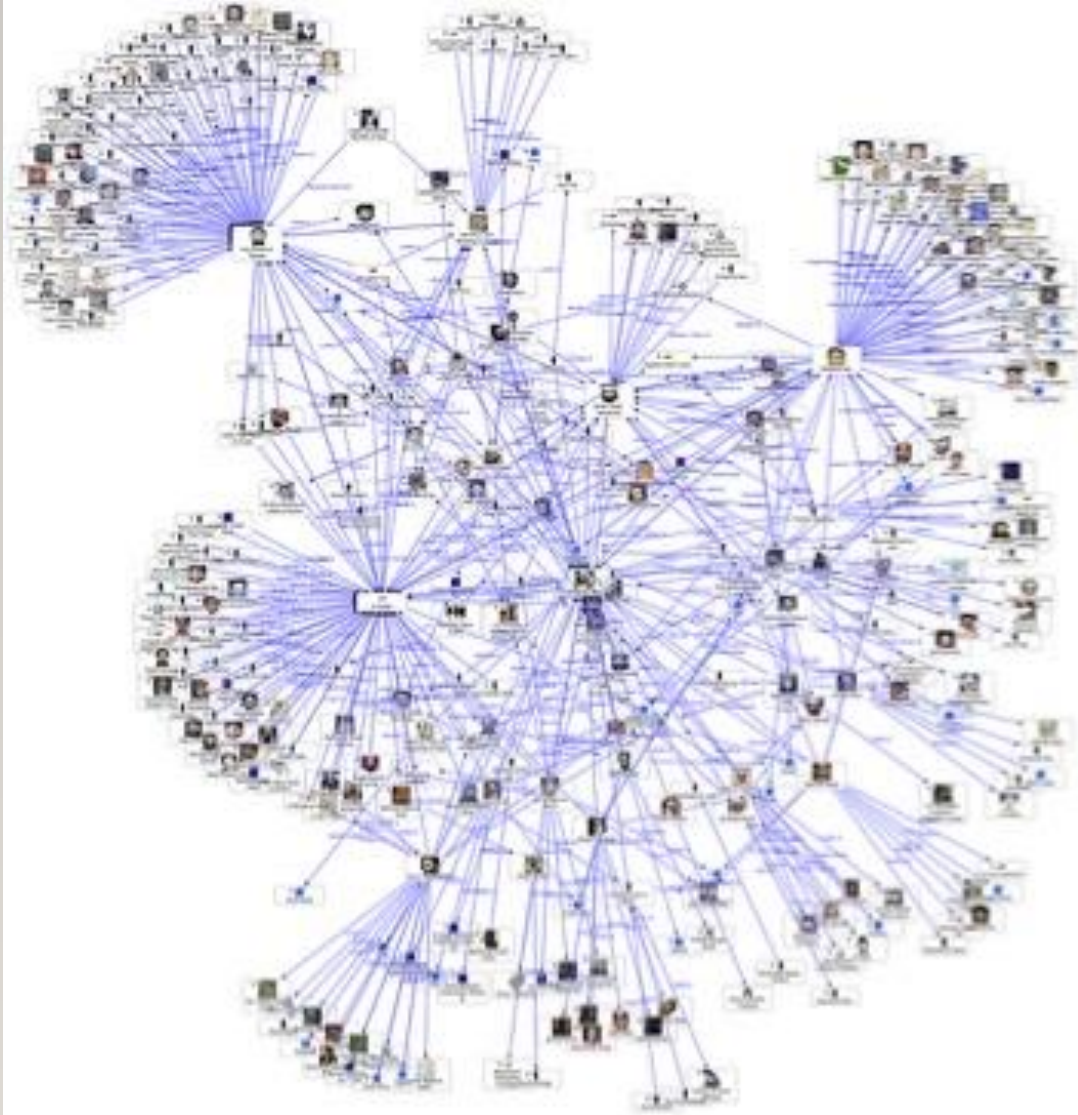
**GRAPH
ANALYTICS**

SOCIAL NETWORK ANALYSIS – INTERPRETING SOCIAL STRUCTURES

- **Definition:** Study of social relationships using network theory.
 - **Nodes:** Individuals, organizations, entities.
 - **Edges:** Social ties.
- **Key concepts:**
 - **Centrality:** Importance of a node (e.g., Degree, Betweenness).
 - **Density:** Overall connectivity of a group.
 - **Path Lengths:** Distances between nodes.
- **Emphasis on interpretation:**
 - Originates from sociology/anthropology.
 - Seeks to interpret network patterns in context (e.g., Person X is an informal leader).
 - Prioritizes accuracy of relationships and rich attribute data.
 - **Traditional scale:** Dozens to thousands of nodes.



GRAPH ANALYTICS – COMPUTING ON NETWORKS AT SCALE



- **Definition:** Algorithmic techniques to analyze data represented as a graph.
 - Vertices & edges as data structures for computation.
- **Scope:** Handles **very large networks** (thousands to billions of nodes).
 - Examples: Web graphs, transportation networks, large social media graphs.
- **Key algorithms:**
 - **Shortest path:** (e.g., Dijkstra's)
 - **Centrality:** (e.g., PageRank, Betweenness computation)
 - **Community detection:** (e.g., Louvain method)
- **Emphasis on scalability & generality:**
 - Focus on efficient computation.
 - Domain-agnostic solutions.
 - Excels at finding patterns quickly.

GRAPH ANALYTICS – UNCERTAIN DEFINITION

COMPUTATION-BASED

- 1) Each query accesses the entire graph
 - 2) Queries iterate until a fixpoint is reached (or other termination condition)
- Examples
 - PageRank
 - All pairs shortest path
 - Connected Components
 - Diameter finding
 - Graph colouring
 - Clustering
 - Pattern mining

OLAP EQUIVALENT

- OLAP over graphs¹
- Supporting analogs of operations such as roll-up, drill-down, slice-and-dice
- Examples (of some work)
 - Graph summarization²
 - Snapshot-based aggregation³
 - Graph cube⁴
 - Pagrol⁵
 - Gagg⁶

¹Gómez et al. Performing OLAP over graph data: query language, implementation, and a case study.” *BIRTE*, 2017.

²Tian, Y., et al. “Efficient aggregation for graph summarization.” *SIGMOD*, 2008.

³Chen, C., et al. “Graph OLAP: Towards online analytical processing on graphs.” *ICDM*, 2008.

⁴Zhao, P., et al. “Graph cube: on warehousing and OLAP multidimensional networks.” *SIGMOD*, 2011.

⁵Wang, Z., et al. “Pagrol: PARallel GRaph OLap over large-scale attributed graphs.”, *ICDE*, 2014.

⁶Maali, F. et al. “Gagg: A graph aggregation operator.” *Semantic Web*, 2015.

SOCIAL NETWORK ANALYSIS VS. GRAPH ANALYTICS

SOCIAL NETWORK ANALYSIS

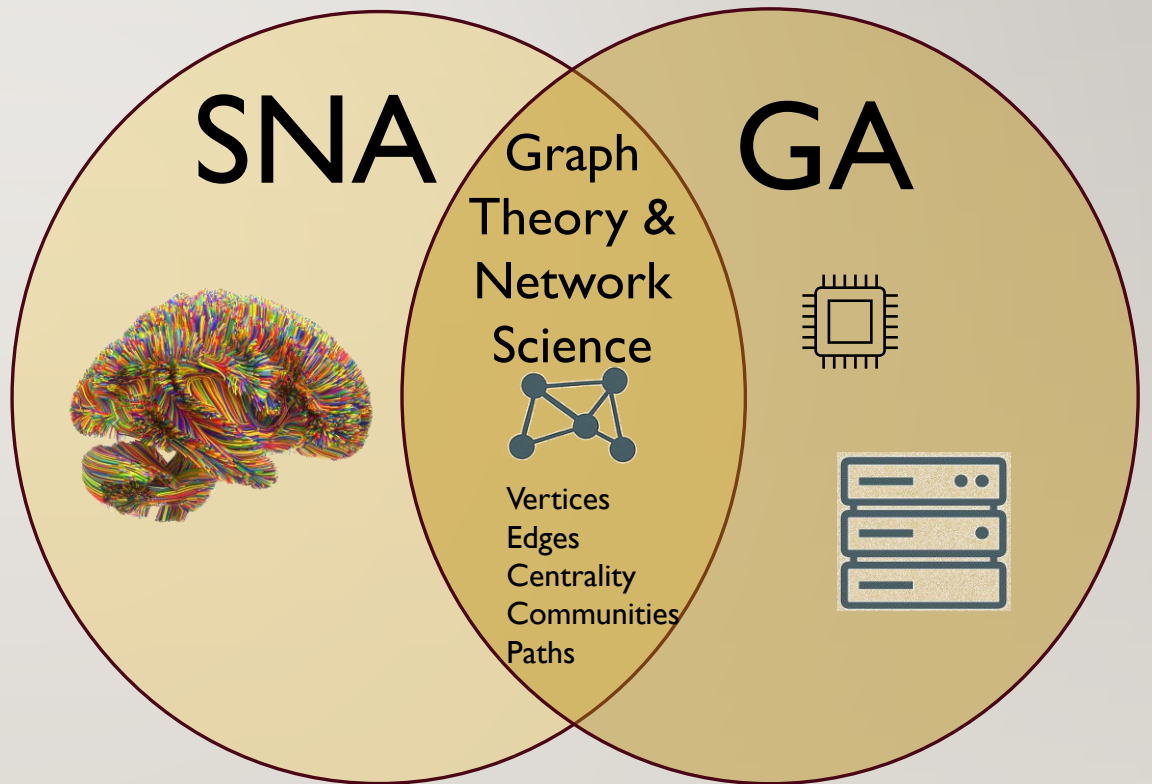
Primary Focus	• Interpretability, social meaning, theory
Typical Scale	
Data Type	• Dozens to thousands of nodes • Curated, survey, rich contextual data
Network Type	• Social networks
Goal	• Explain <i>why</i> patterns exist, understand social mechanisms

GRAPH ANALYTICS

- Scalability, algorithms, computation
- Thousands to billions of vertices
- Observational, machine-collected, large-scale
- Network-agnostic
- Find patterns efficiently, quantify properties, optimize tasks

COMMON GROUND: GRAPH THEORY & NETWORK CONCEPTS

- **Shared foundation:** Both rely on fundamental **Graph Theory**.
 - Vertices, Edges, Paths, Degrees, etc. are universal.
- **Overlapping metrics & algorithms:**
 - **Centrality:** Originated in SNA, computed at scale by GA.
 - **Community detection:** SNA concept of cohesive subgroups, GA provides algorithms (e.g., Louvain).
- **Interdisciplinary network science:**
 - Emergence of a unified field.
 - Discoveries like small-world and scale-free networks blend both.
- **Collaboration potential:** SNA concepts inform GA algorithms; GA results interpreted by SNA theories.



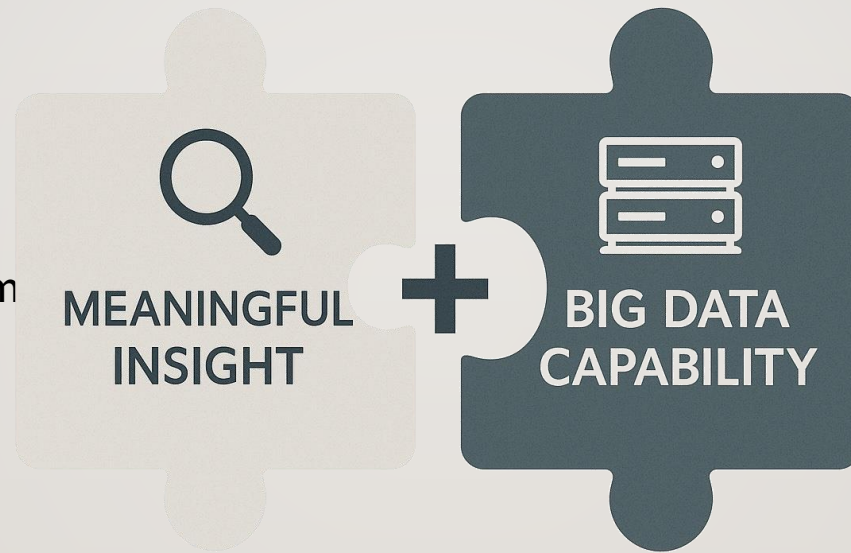
WHY INTEGRATION? – BRIDGING INSIGHT AND SCALE

- **Limitations in isolation:**

- **SNA alone:** Overwhelmed by massive data (e.g., social media platforms).
- **GA alone:** May find patterns but struggle to explain or leverage them meaningfully.

- **The Synergy:**

- GA handles **scale and complexity**.
- SNA guides **analysis** and **interprets results** in human terms.



- **GA augments SNA:**

- Enables analysis of much larger, more complex networks.
- Tests SNA hypotheses on big data.

- **SNA enriches GA:**

- Ensures results are **interpretable and actionable**.
- Helps set the right objectives for graph analytics.

AGENDA

Introduction & Foundations

- Why connect SNA & GA?
- Defining SNA & GA
- Common ground & why integrate?

Five Examples

Graph Analytics Platforms

Integration, Tools & Methodologies

Challenges & Future Directions

AGENDA

Introduction & Foundations

Five Examples

- Influence
- Communities
- Diffusion
- Link prediction
- Roles

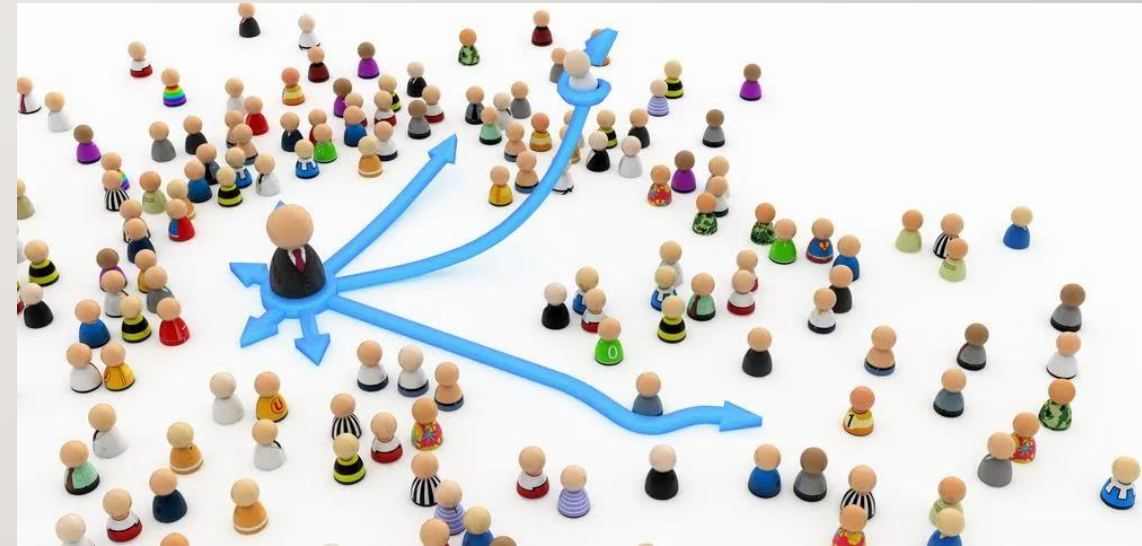
Graph Analytics Platforms

Integration, Tools & Methodologies

Challenges & Future Directions

INFLUENCE IN NETWORKS – WHO MATTERS?

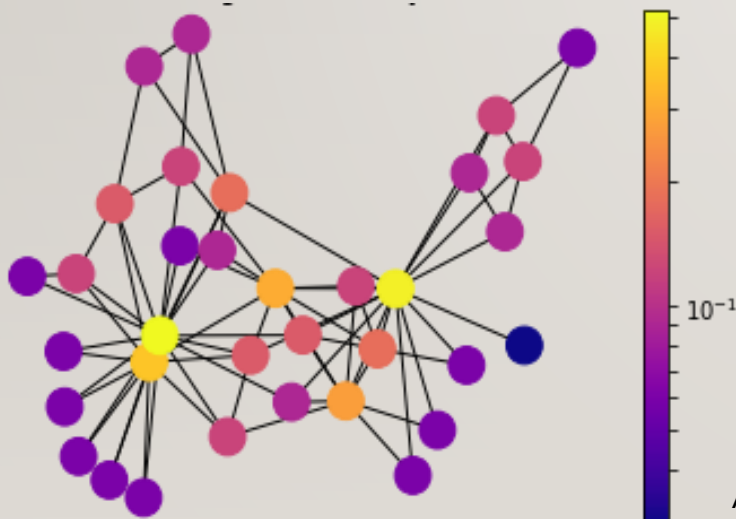
- **What is influence?**
 - Ability to affect others (spread info, exert opinion).
 - Can be well-connected or strategically positioned.
- **Why it matters:**
 - **Marketing:** Target viral spreaders.
 - **Organizations:** Find informal leaders, bottlenecks.
 - **Public Health:** Identify super-spreaders.
- **SNA perspective:**
 - Historically studied through **Centrality** measures.
 - Qualitative understanding of actor's network position.
- **GA challenge:**
 - Efficient computation on **large networks**.
 - Algorithms for **Influence Maximization**.



MEASURING INFLUENCE: CENTRALITY METRICS (SNA VIEW)

DEGREE CENTRALITY

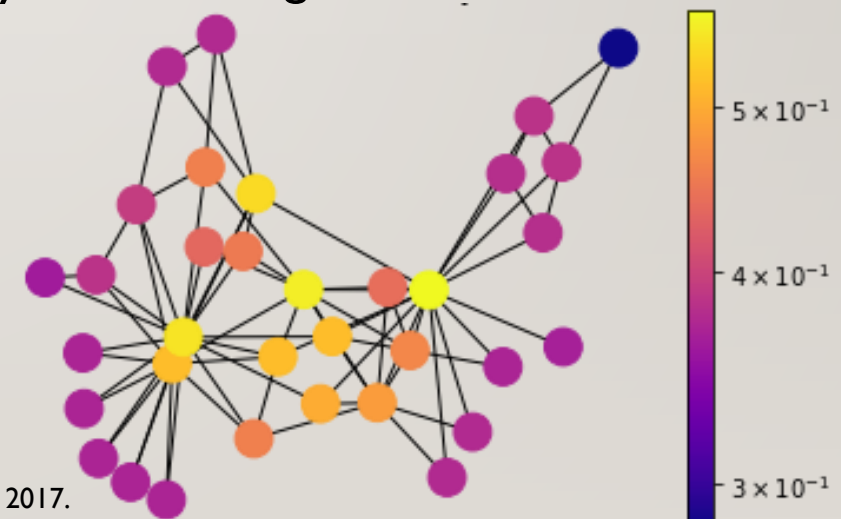
- **Definition:** Number of direct connections a node has.
- **Interpretation:** Visibility, popularity, immediate reach.
- *Example:* A person with many friends on social media.



Aksakalli, C.G.. Network Centrality Measures, 2017.
<http://bit.ly/46fEQqs>

CLOSENESS CENTRALITY

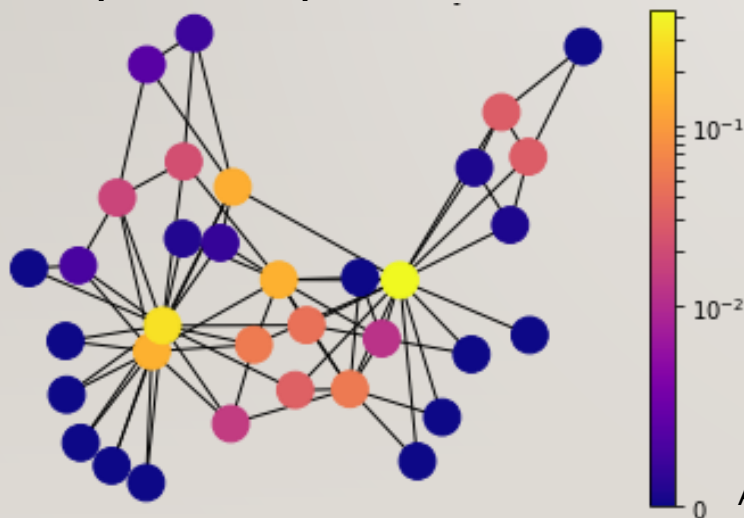
- **Definition:** Average shortest path distance to all other nodes.
- **Interpretation:** Efficiency in spreading information, independence from intermediaries.
- *Example:* A person who can quickly reach anyone in an organization.



MEASURING INFLUENCE: CENTRALITY METRICS (SNA VIEW)

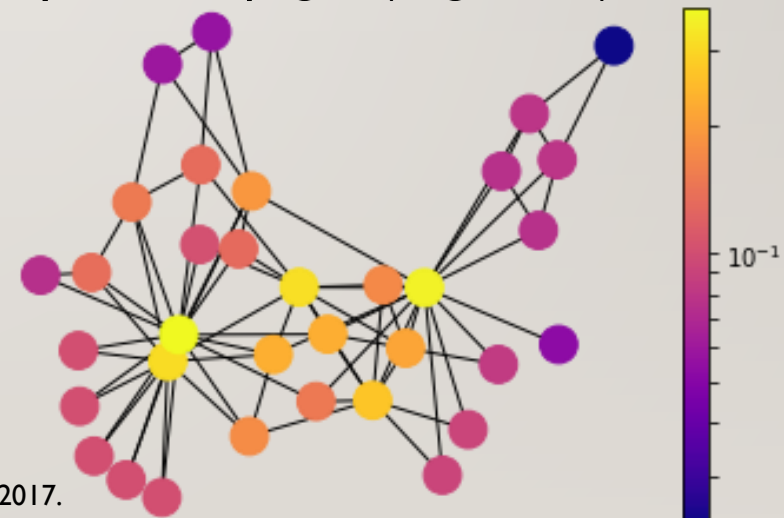
BETWEENNESS CENTRALITY

- **Definition:** How often a node lies on the shortest paths between other pairs of nodes.
- **Interpretation:** Brokerage, gatekeeping, control over information flow.
- *Example:* A person connecting two otherwise separate departments.



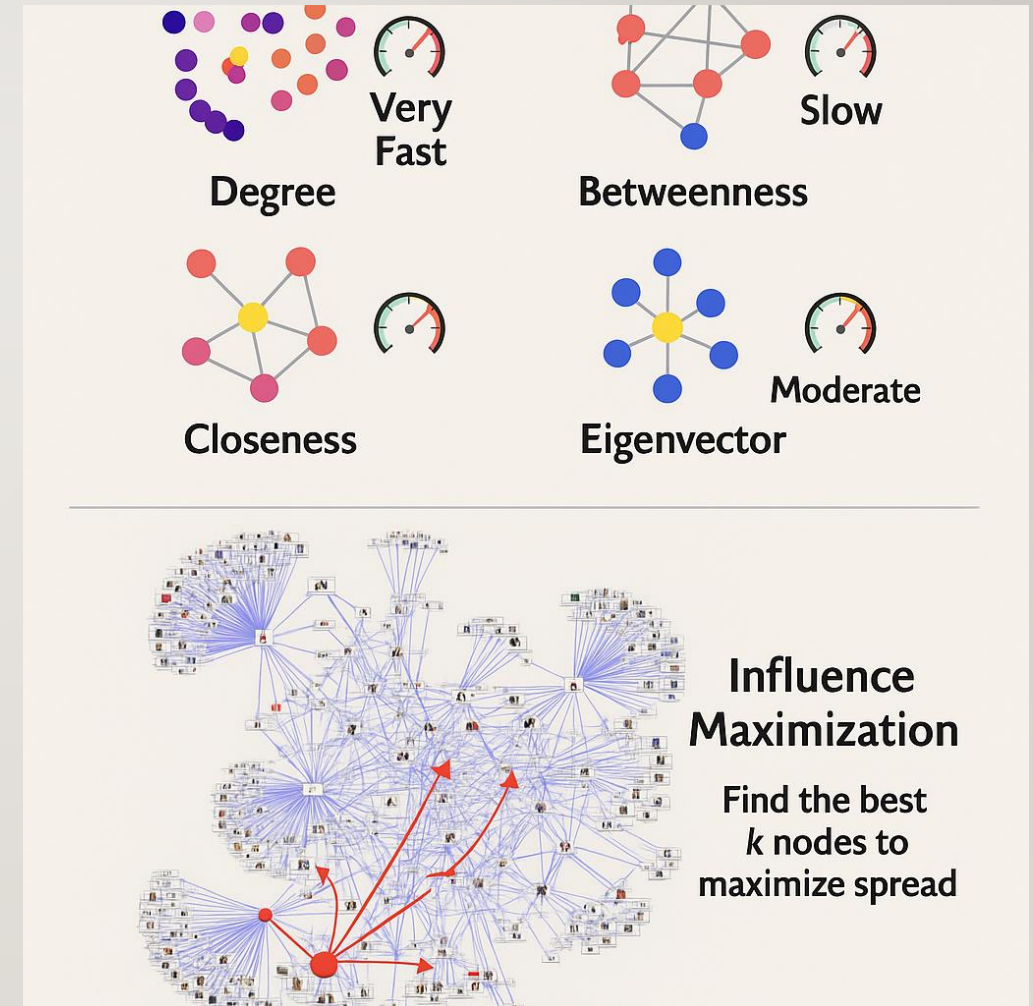
EIGENVECTOR CENTRALITY

- **Definition:** A node is important if it's connected to other important nodes.
- **Interpretation:** Prestige, influence by association, being part of an influential core.
- *Example:* A web page linked by many other highly-ranked pages (PageRank).



GA FOR INFLUENCE: COMPUTATION & MAXIMIZATION

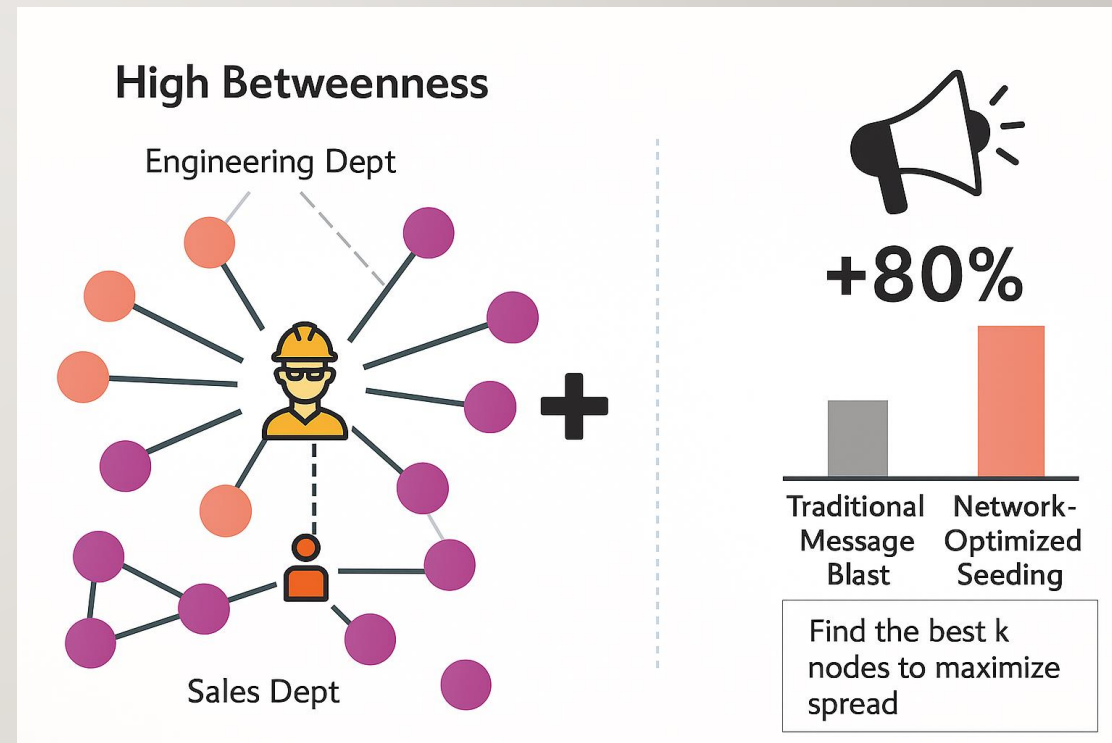
- **Computing centrality at scale:**
 - **Degree:** Trivial for massive graphs.
 - **Closeness:** Feasible with approximations or parallel computing.
 - **Betweenness:** Computationally heavy, but algorithms (e.g., Brandes') make it feasible for large graphs.
 - **Eigenvector/PageRank:** Highly scalable, iterative algorithms (e.g., power iteration for billions of vertices).
- **Beyond centrality: Influence maximization¹:**
 - **Problem:** Given a budget of k nodes, which k should be chosen to maximize influence spread?
 - **Approach:** Uses diffusion simulations and optimization algorithms (e.g., greedy algorithms with submodular optimization).
 - **Example:** Viral marketing – picking optimal “seed” individuals.



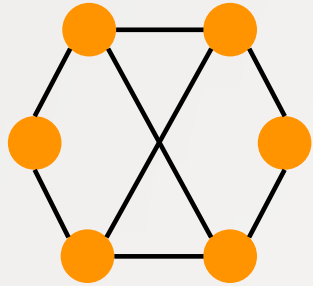
¹Kempe, D., et al. "Maximizing the spread of influence through a social network." KDD, 2003.

EXAMPLE: FINDING INFLUENCERS IN A SOCIAL NETWORK

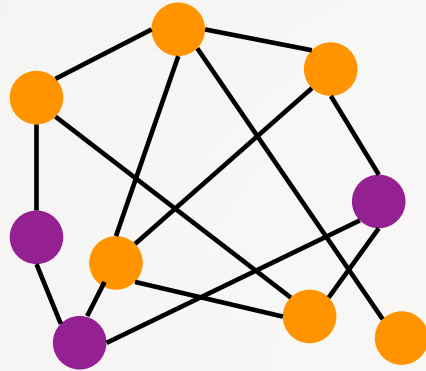
- **Scenario: corporate communication network**
 - **Network:** Company email communication (vertices = employees, edges = frequent emails).
 - **GA:** Computed Degree & Betweenness centrality, and ran Influence Maximization simulations.
 - **SNA interpretation:** Validated algorithmic findings with employee roles and context.
- **Key Insights:**
 - **Hidden bridges:** An engineer, not a manager, had *highest betweenness centrality*.
 - *SNA Insight:* Recognized as a critical informal liaison between departments.
 - **Diverse reach:** Influence maximization suggested seeding messages via cross-team connectors.
 - *SNA Insight:* Achieved 90% company reach with only 5 seeds, outperforming hierarchical blast.
 - **Actionable knowledge:** Guided targeted internal communication strategies.



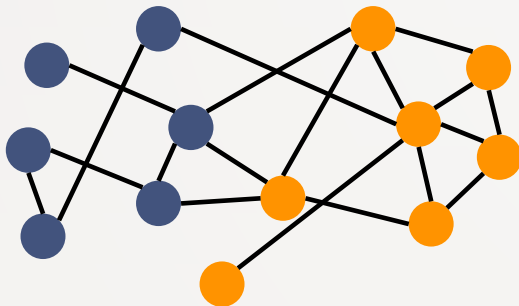
SOCIAL NETWORK ANALYSIS OF COMMUNITIES – COHESIVE SUBGROUPS



Clique
(Ideal cohesion)



High Density
(e.g., over 80% connected)

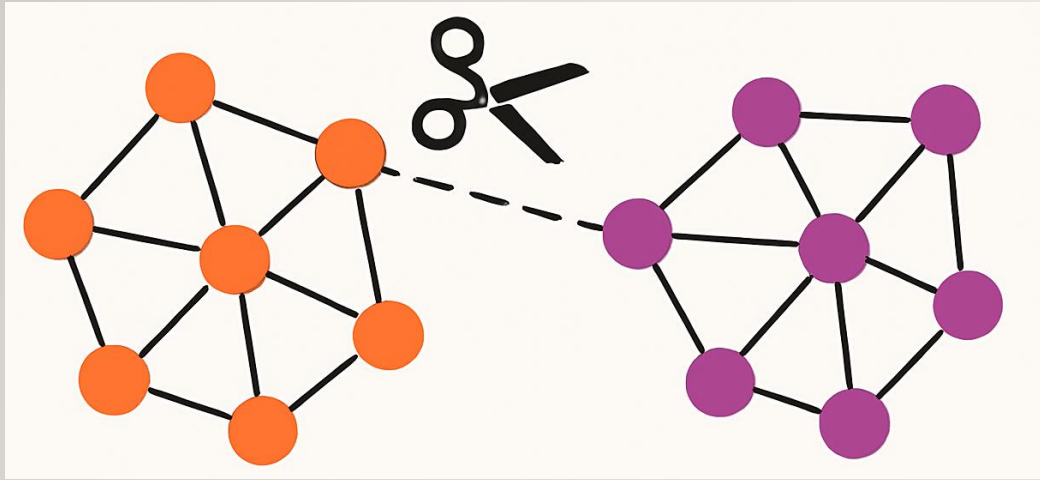


Zachary's Karate Club¹

- **Defining cohesion:**
 - **Cliques:** All nodes directly connected (ideal, rare).
 - **Near-cliques / k-cores:** Relaxed conditions for strong groups.
 - **Density:** Proportion of actual ties to possible ties within a group.
 - *High density = strong community.*
- **Identifying communities in SNA:**
 - Historically, visual inspection or small-scale techniques.
 - Relied on node attributes (e.g., department, shared interest).
- **Emphasis on interpretation:**
 - What common characteristic or context binds these people?
 - *Example:* Zachary's Karate Club – structural split matched real-world schism.
 - Translate abstract clusters into meaningful labels (e.g., K-Pop Fans Community).

¹Zachary. "An information flow model for conflict and fission in small groups. " J. Anthropological Research, 1977.

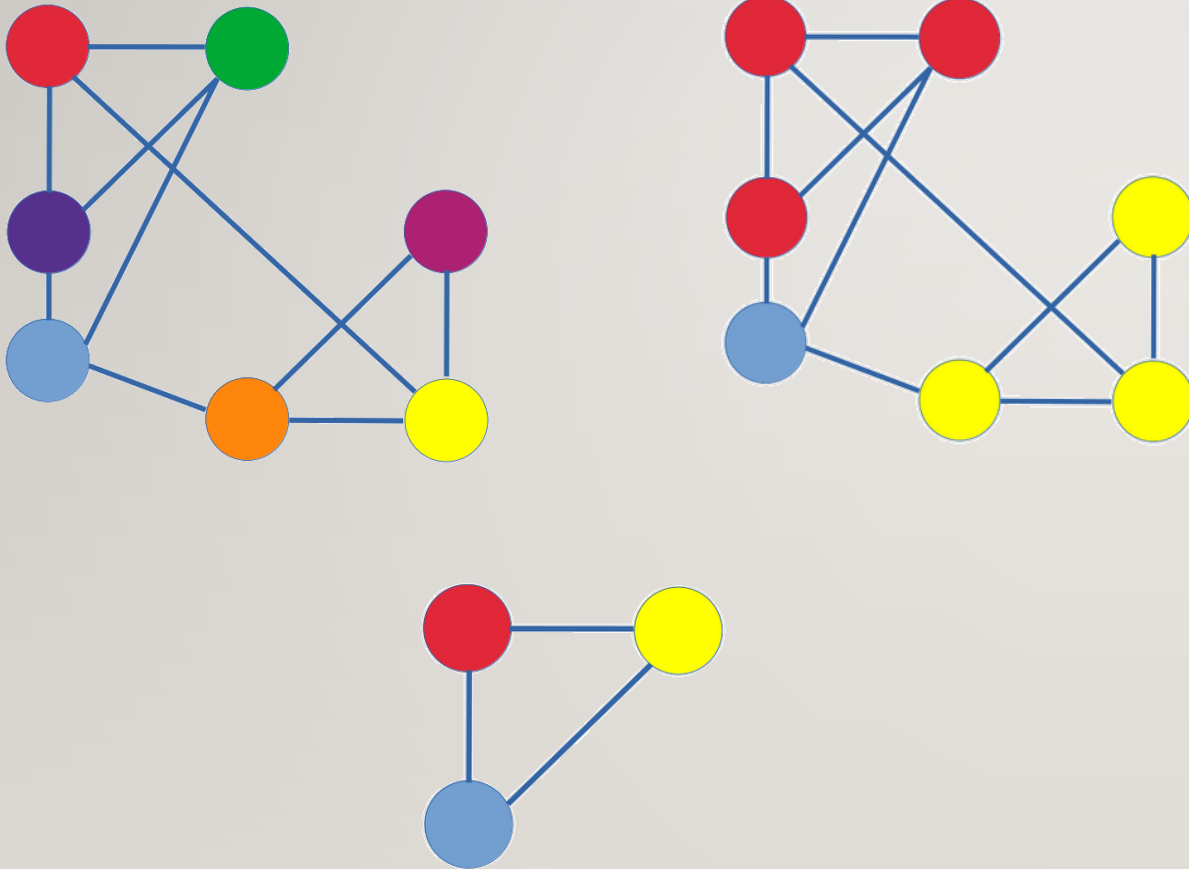
GRAPH ANALYTICS FOR COMMUNITY DETECTION



- **Algorithmic approaches:**
 - **Girvan–Newman¹:** Iteratively removes bridge edges with high betweenness.
 - *Concept:* Cutting weak links between strong groups.
 - *Output:* Dendrogram of hierarchical communities.

¹ Girvan, M. & Newman, M.E.J. “Community structure in social and biological networks.” *PNAS*, 2002.

GRAPH ANALYTICS FOR COMMUNITY DETECTION



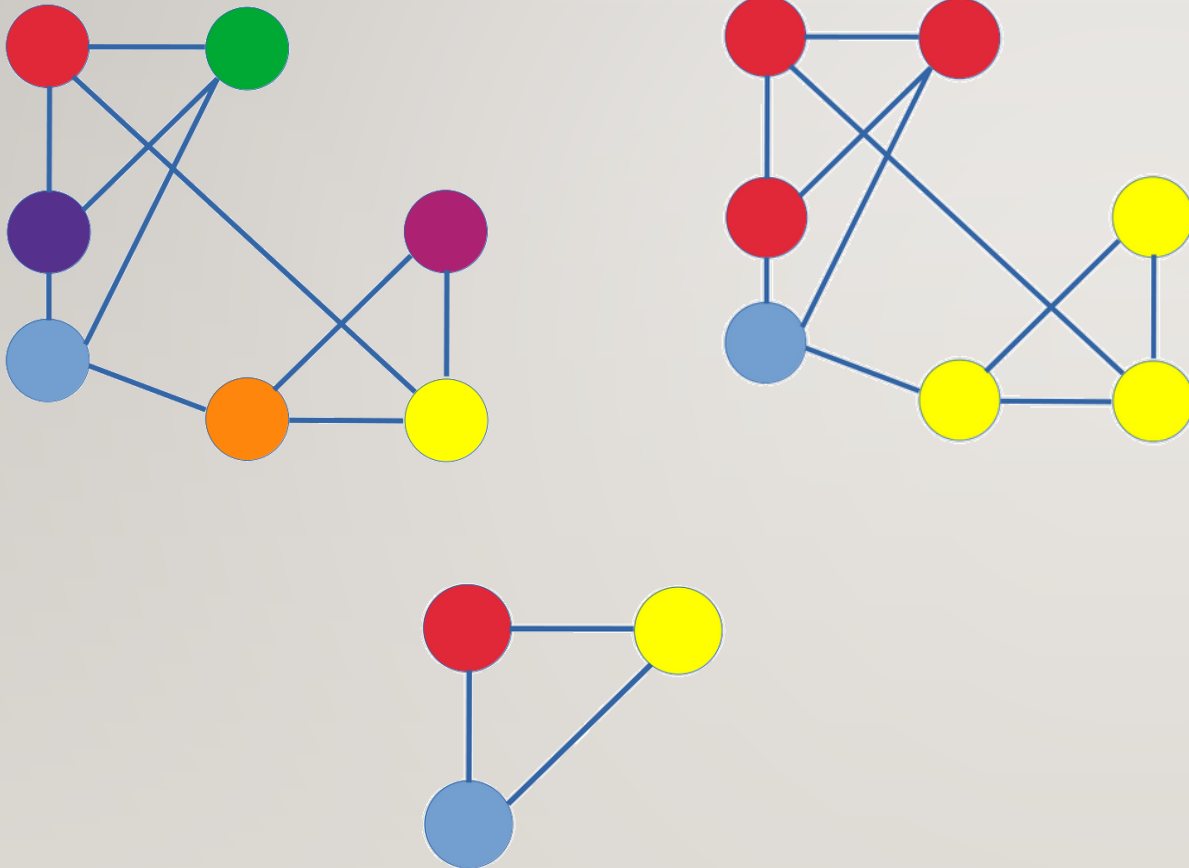
- **Algorithmic approaches:**

- **Girvan–Newman¹:** Iteratively removes bridge edges with high betweenness.
 - *Concept:* Cutting weak links between strong groups.
 - *Output:* Dendrogram of hierarchical communities.
- **Modularity optimization (e.g., Louvain²):**
 - **Modularity:** Measures density of internal vs. external connections (higher = stronger structure).
 - **Louvain:** Greedy algorithm, merges communities to maximize modularity.
 - *Advantage:* Fast and scalable for large networks (millions of nodes).

¹ Girvan, M. & Newman, M.E.J. "Community structure in social and biological networks." *PNAS*, 2002.

² Blondel, V D et al. "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

GRAPH ANALYTICS FOR COMMUNITY DETECTION



- **Algorithmic approaches:**

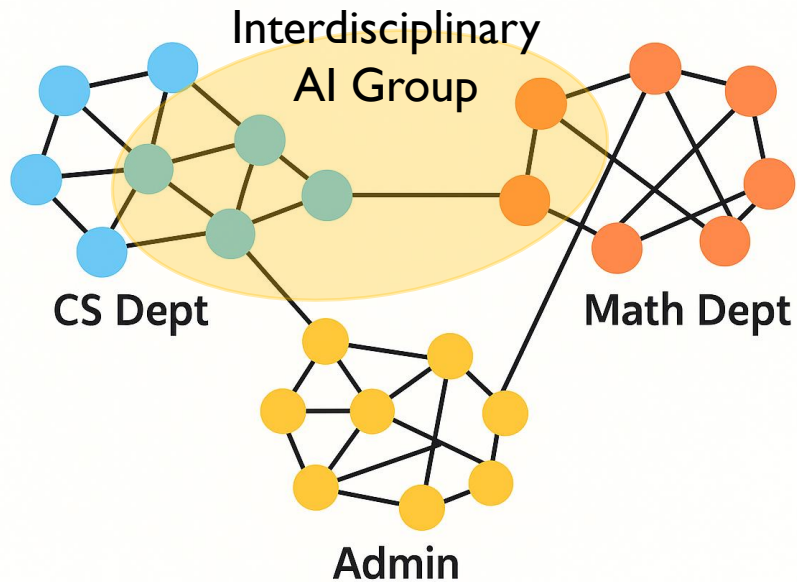
- **Girvan–Newman¹:** Iteratively removes bridge edges with high betweenness.
 - *Concept:* Cutting weak links between strong groups.
 - *Output:* Dendrogram of hierarchical communities.
- **Modularity optimization (e.g., Louvain²):**
 - **Modularity:** Measures density of internal vs. external connections (higher = stronger structure).
 - **Louvain:** Greedy algorithm, merges communities to maximize modularity.
 - *Advantage:* Fast and scalable for large networks (millions of nodes).

- **Scalability:** Process massive networks quickly.
- **Output:** Groupings (community labels for each node).
- **Post-processing:** Requires SNA interpretation to name and understand.

¹ Girvan, M. & Newman, M.E.J. "Community structure in social and biological networks." *PNAS*, 2002.

² Blondel, V D et al. "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

EXAMPLE: COMMUNITY DETECTION IN A REAL NETWORK



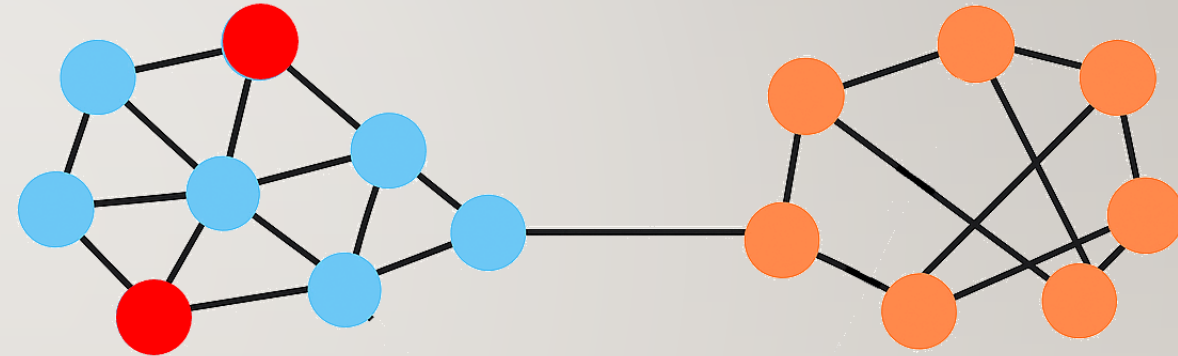
- **Scenario: University email communication network**
 - **Network:** Email interactions among university faculty, staff, and students.
 - **GA:** Applied Louvain method to detect communities.
 - **SNA Interpretation:** Examined node affiliations within detected clusters.

Key Insights:

- **Rediscovering structure:** Major communities strongly corresponded to **academic departments** (e.g., CS, Math, Admin).
 - *Validation:* Algorithm effectively rediscovered organizational structure from communication patterns.
- **Hidden collaborations:** Detected a smaller, cross-departmental community.
 - *Insight:* An interdisciplinary AI research group, not obvious from formal structure.
- **Actionable knowledge:** Quantified inter-departmental communication, revealing potential silos.

DIFFUSION IN NETWORKS – HOW THINGS SPREAD

- **What is diffusion?**
 - Process where something (info, behaviour, disease) spreads along network edges.
 - Can be **contagious** or **influence-based**.
- **Importance:**
 - Predict virality (marketing, news).
 - Forecast epidemic outcomes (public health).
 - Manage cascade effects (misinformation).
- **SNA perspective:**
 - Theoretical models: Threshold Models, Epidemic Models (SIR).
 - **Strength of Weak Ties**¹: Weak ties (bridges) are crucial for widespread.
- **GA perspective:**
 - **Simulation:** Model spread on large networks (what if scenarios).
 - **Influence spread calculation:** Evaluate reach of seed sets.
 - **Identifying critical nodes/links:** Super-spreaders, bottlenecks.

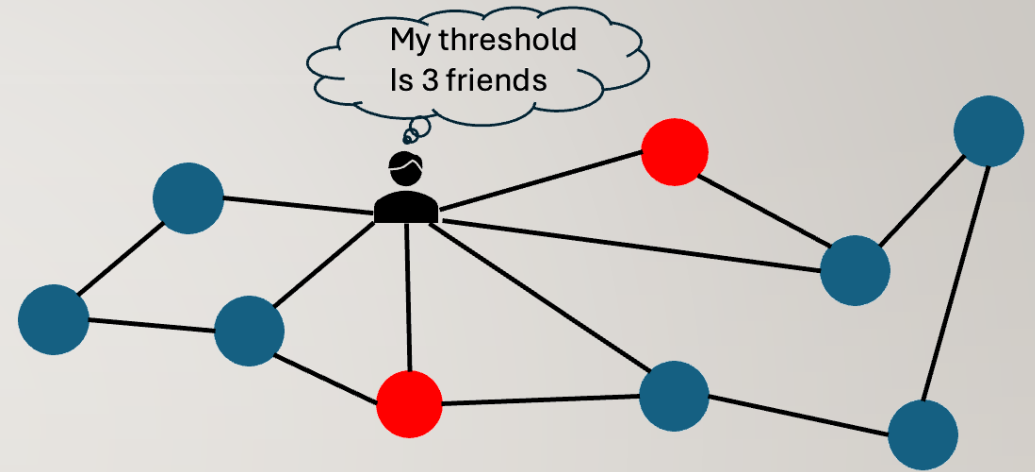


¹Granovetter, M.S. "The Strength of Weak Ties." *Am. J. Sociology*, 1973.

SNA DIFFUSION MODELS – THRESHOLDS & CONTAGION

- **Threshold models (complex contagion):**

- **Concept:** adopts only after a *certain fraction/number* of neighbours adopt.
- **Mechanism:** Captures social reinforcement, peer pressure, critical mass.
- *Example:* Joining a social movement if enough friends do.



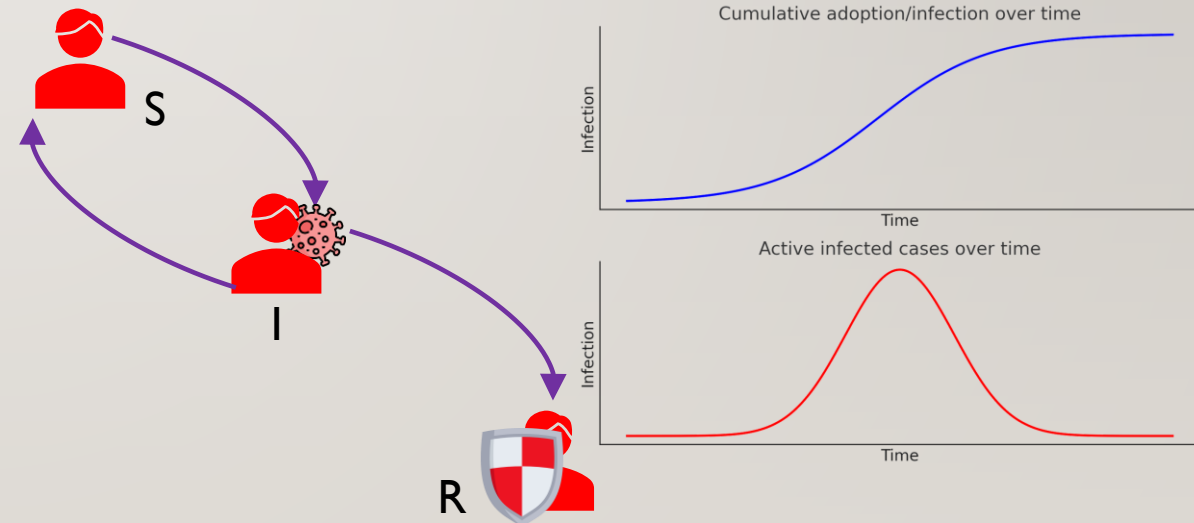
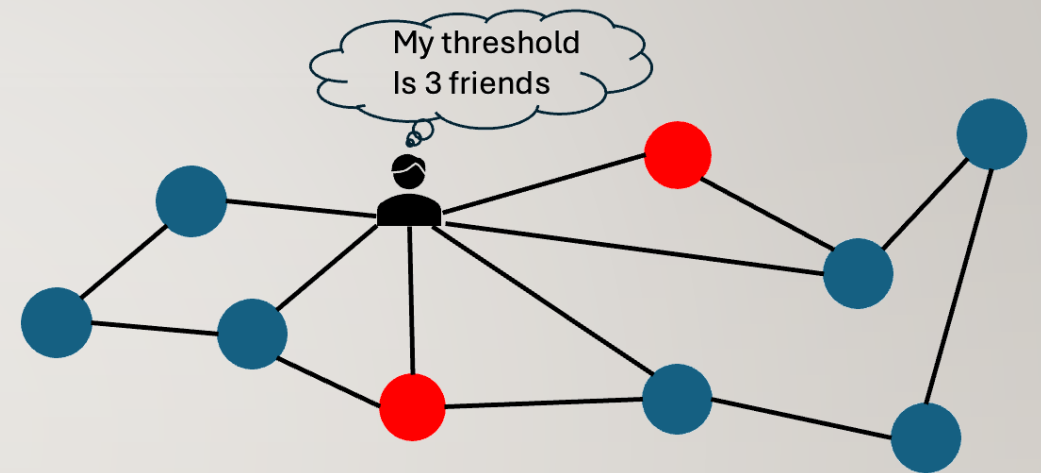
SNA DIFFUSION MODELS – THRESHOLDS & CONTAGION

- **Threshold models (complex contagion):**

- **Concept:** adopts only after a *certain fraction/number* of neighbours adopt.
- **Mechanism:** Captures social reinforcement, peer pressure, critical mass.
- *Example:* Joining a social movement if enough friends do.

- **Epidemic models (Simple Contagion):**

- **Concept:** One exposure can be enough (probabilistic spread).
- **Models¹:**
 - **SI (Susceptible-Infected):** Infection spreads indefinitely.
 - **SIR (Susceptible-Infected-Recovered):** Nodes gain immunity.
 - **SIS (Susceptible-Infected-Susceptible):** No immunity, can lead to endemic state.



¹Kermack, W.O. & McKendrick, A. G. .“A Contribution to the Mathematical Theory of Epidemics.” *Proc. Royal Society A*, 1927.

SNA DIFFUSION MODELS – THRESHOLDS & CONTAGION

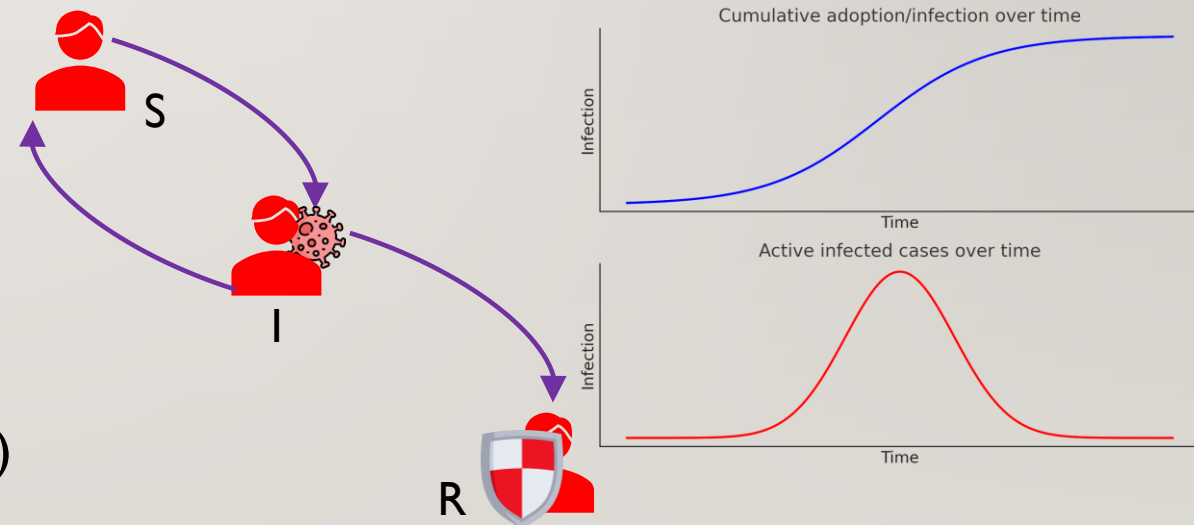
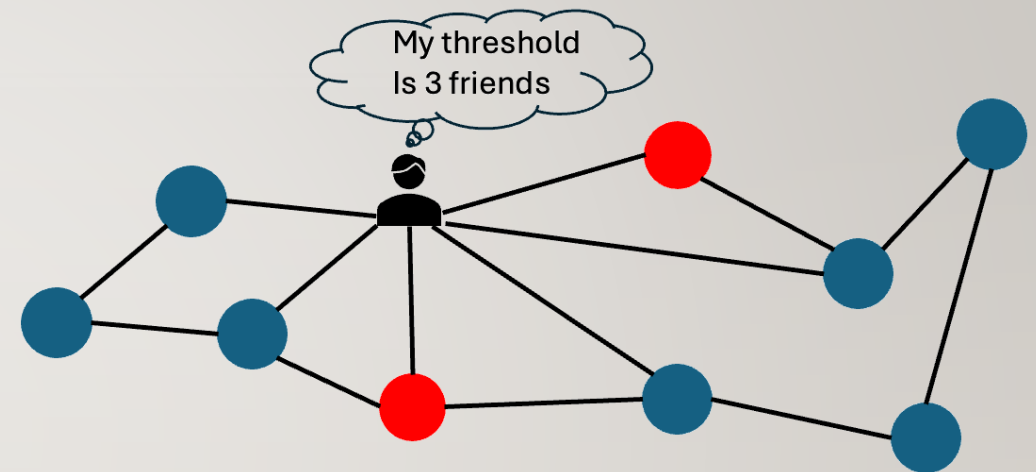
- **Threshold models (complex contagion):**

- **Concept:** adopts only after a *certain fraction/number* of neighbours adopt.
- **Mechanism:** Captures social reinforcement, peer pressure, critical mass.
- *Example:* Joining a social movement if enough friends do.

- **Epidemic models (Simple Contagion):**

- **Concept:** One exposure can be enough (probabilistic spread).
- **Models¹:**
 - **SI (Susceptible-Infected):** Infection spreads indefinitely.
 - **SIR (Susceptible-Infected-Recovered):** Nodes gain immunity.
 - **SIS (Susceptible-Infected-Susceptible):** No immunity, can lead to endemic state.

- **Key insight:** Network structure (hubs, short paths) significantly impacts spread.



¹Kermack, W.O. & McKendrick, A. G. "A Contribution to the Mathematical Theory of Epidemics." *Proc. Royal Society A*, 1927.

GA FOR DIFFUSION: SIMULATION & ANALYSIS

- **Simulation of spread:**

- Model probabilistic or threshold-based propagation on large networks.
- Use: What if scenarios, parameter sweeps.

- **Influence spread calculation:**

- Core of *Influence Maximization* (evaluating seed sets).
- **Key Insight:** Network structure (hubs, short paths) significantly impacts spread.

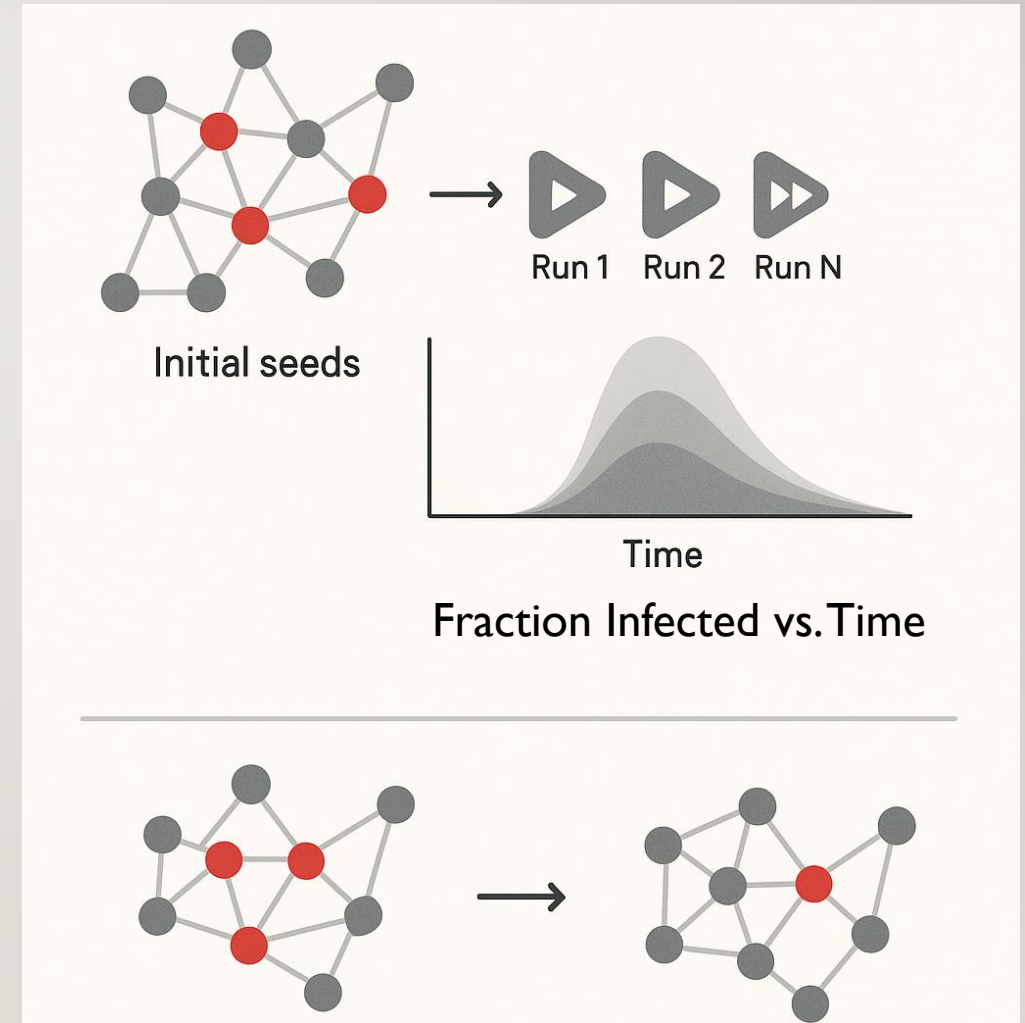
- **Analytical approximations:**

- Estimate epidemic thresholds (e.g., based on network eigenvalues).

- **Identifying critical nodes/links:**

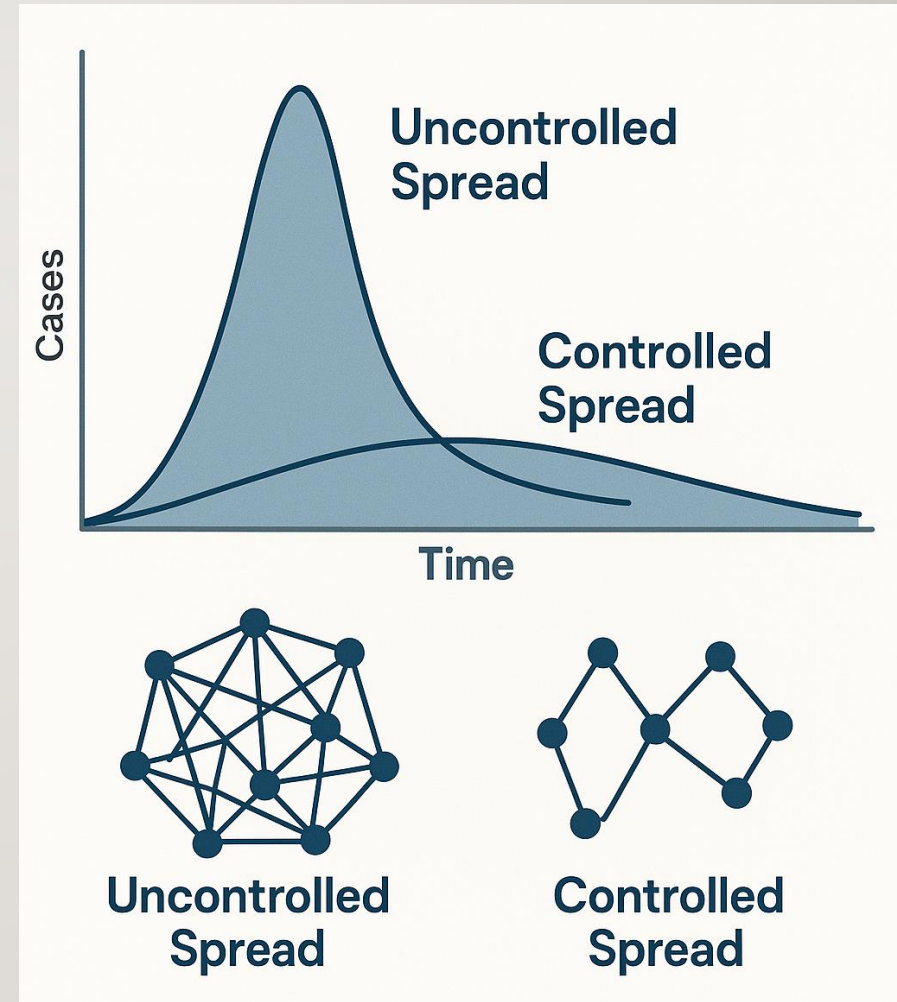
- Leverage centrality measures (e.g., betweenness for super-spreaders).
- Algorithms to find minimal sets for network dismantling.

- **Large-scale applications:** City-level epidemic modelling, predicting content virality.

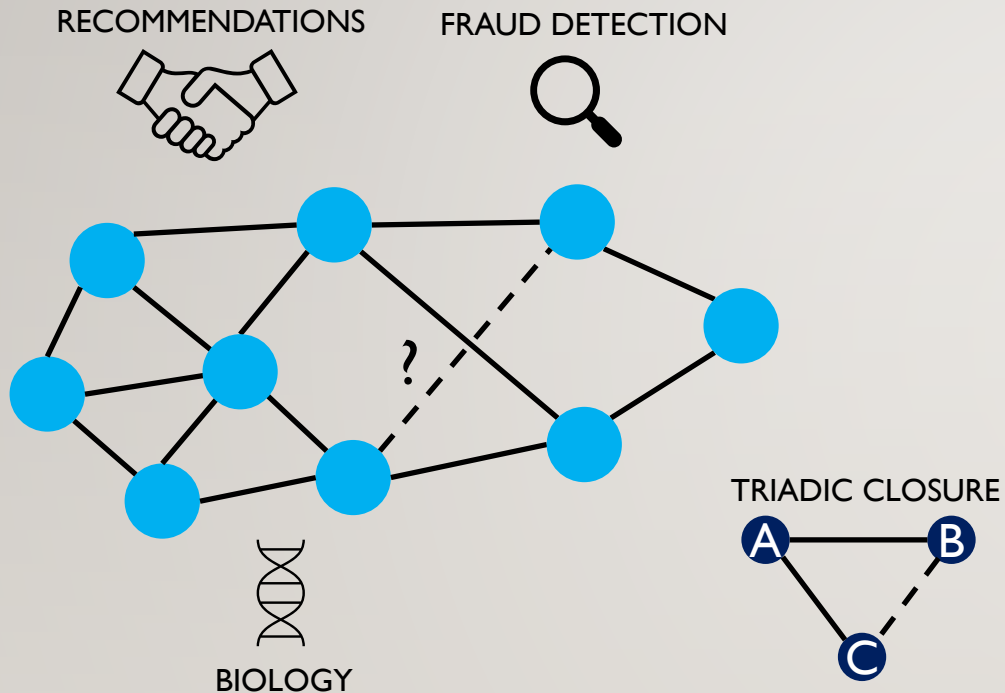


EXAMPLE: MODELING SPREAD IN A NETWORK

- **Scenario: COVID-19 outbreak in a city contact network**
 - **Network model:** Individuals as nodes, frequent close contacts as edges.
 - **GA:** SIR model simulations to project the outbreak under policies.
 - **SNA interpretation:** Connected model outcomes to social behaviours.
- **Key insights:**
 - **Policy impact:** Simulation predicted *flattening the curve* with interventions (e.g., school closures).
 - *SNA Insight:* Closing schools cut many inter-family edges, confirming children as super-spreaders.
 - **Targeted protection:** Hub individuals (many daily contacts) were consistently infected in simulations.
 - *Insight:* Highlighted need to protect specific 'role' types (e.g., transit workers).
 - **Actionable strategy:** Informed public health policymakers on effective interventions.



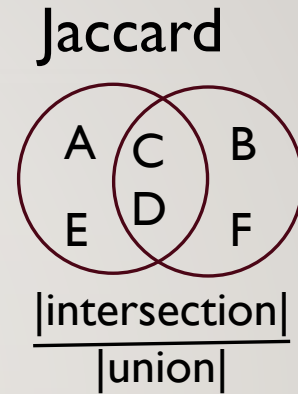
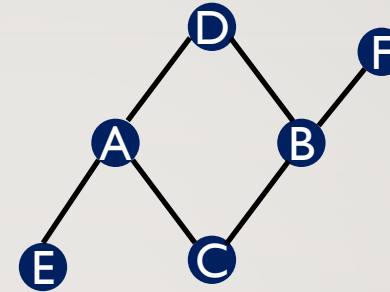
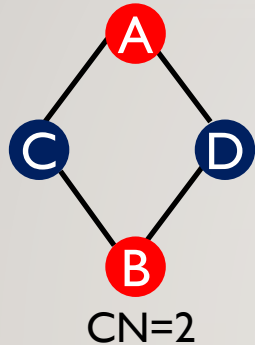
LINK PREDICTION – ANTICIPATING NETWORK EVOLUTION



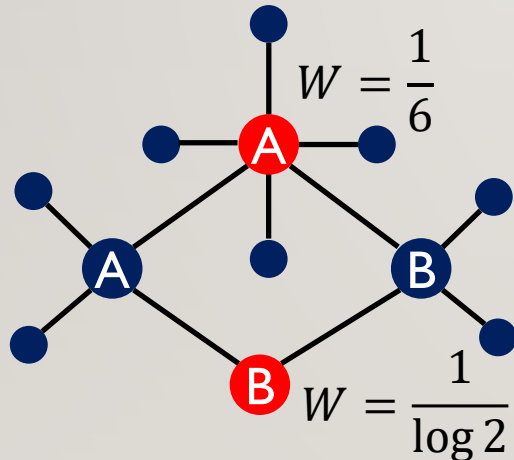
- Inferring which non-connected node pairs will likely form a connection in the future; or
- Detecting currently unobserved (hidden) connections.
- **Importance:**
 - **Recommendations:** People You May Know.
 - **Network growth analysis:** Understanding evolution mechanisms.
 - **Fraud detection:** Identifying hidden conspiratorial links.
 - **Biology:** Predicting protein interactions, gene links.
- **SNA Perspective:**
 - Theories on tie formation: *Homophily* (similarity), *Triadic Closure* (friend-of-a-friend).
 - **Opportunity/Propinquity:** Proximity increases likelihood.
- **GA Perspective:**
 - Quantifies similarity (common neighbours, structural position).
 - Uses ML algorithms to combine factors and learn from past growth.
 - Scans millions of pairs efficiently.

HEURISTICS FOR LINK PREDICTION – COMMON NEIGHBOURS & SIMILARITY

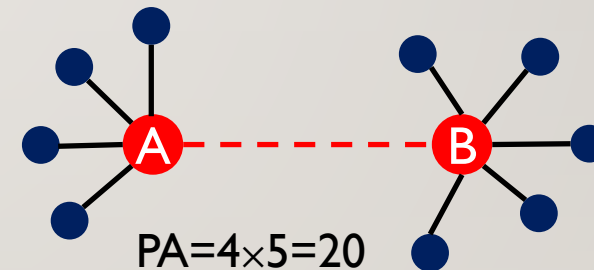
Common Neighbours



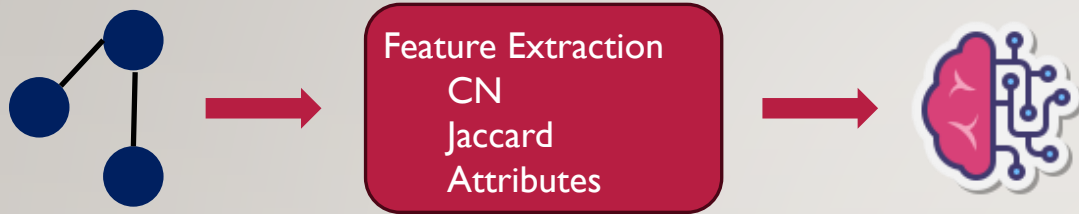
Ademic-Adar



Preferential Attachment



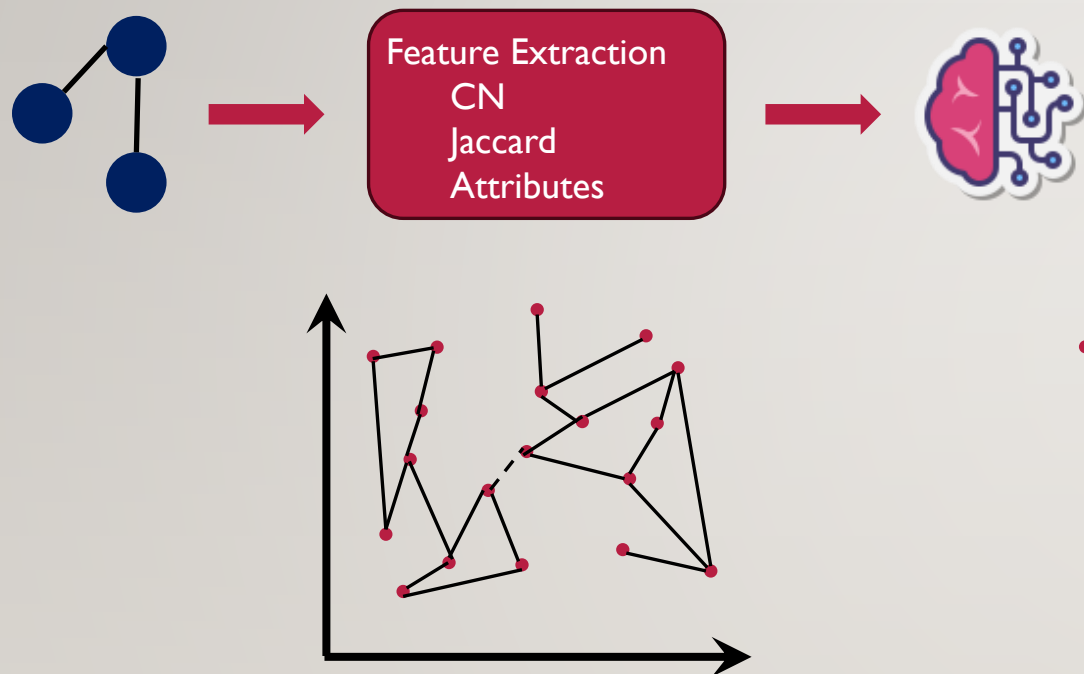
BEYOND HEURISTICS: MACHINE LEARNING & GRAPH EMBEDDINGS



- **Feature-based machine learning¹:**
 - Assemble feature vectors for vertex pairs (e.g., common neighbours, attribute similarity, degree differences).
 - Train classifiers on historical link formation.
 - **Advantage:** Captures complex, non-linear relationships between features.

¹Backstrom, L. & Leskovec, J. "Supervised Random Walks: Predicting and Recommending Links." WSDM, 2011.

BEYOND HEURISTICS: MACHINE LEARNING & GRAPH EMBEDDINGS

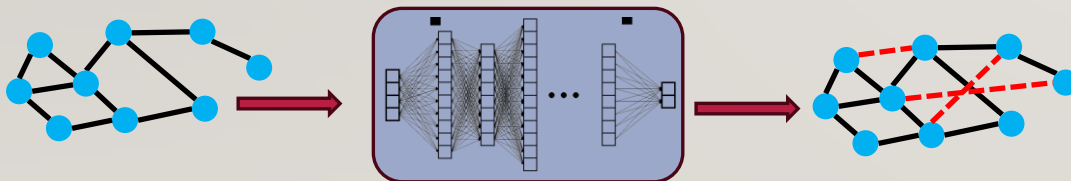
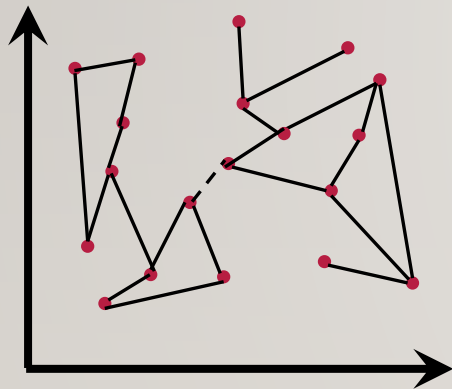
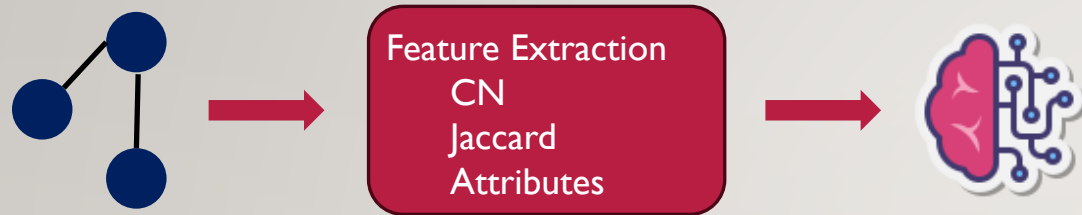


- **Feature-based machine learning¹:**
 - Assemble feature vectors for vertex pairs (e.g., common neighbours, attribute similarity, degree differences).
 - Train classifiers on historical link formation.
 - **Advantage:** Captures complex, non-linear relationships between features.
- **Graph embeddings²:**
 - Generate vector representations for each vertex.
 - **Prediction:** Link likelihood by vector similarity.
 - **Advantage:** Implicitly combines many structural features, scalable.

¹Backstrom, L. & Leskovec, J. "Supervised Random Walks: Predicting and Recommending Links." WSDM, 2011.

²Perrozi, B. et al. "DeepWalk: Online Learning of Social Representations." SIGMOD, 2014.

BEYOND HEURISTICS: MACHINE LEARNING & GRAPH EMBEDDINGS



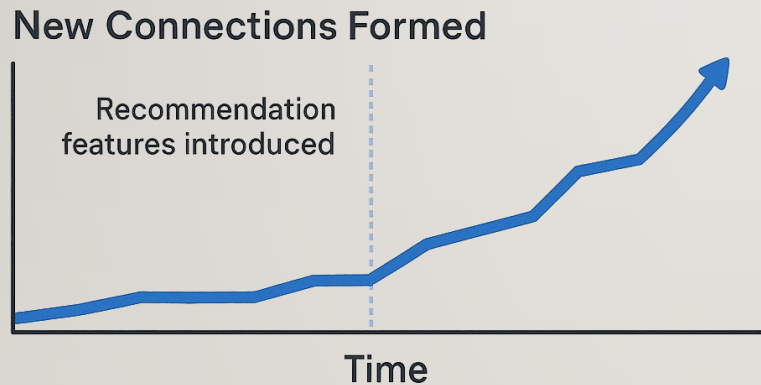
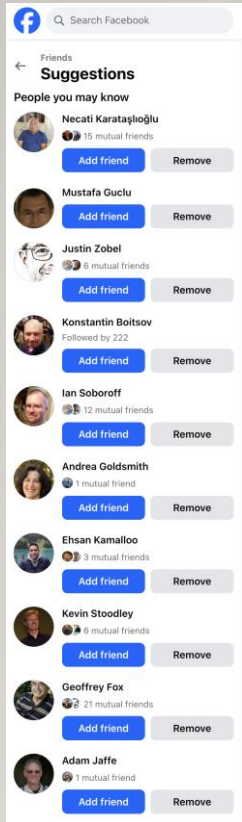
- **Feature-based machine learning¹:**
 - Assemble feature vectors for vertex pairs (e.g., common neighbours, attribute similarity, degree differences).
 - Train classifiers on historical link formation.
 - **Advantage:** Captures complex, non-linear relationships between features.
- **Graph embeddings²:**
 - Generate vector representations for each vertex.
 - **Prediction:** Link likelihood by vector similarity.
 - **Advantage:** Implicitly combines many structural features, scalable.
- **Graph Neural Networks (GNNs)³:**
 - Learn vertex representations and predict links directly.
 - **Advantage:** Learn complex, higher-order patterns, integrate structure and attributes.

¹Backstrom, L. & Leskovec, J. "Supervised Random Walks: Predicting and Recommending Links." WSDM, 2011.

²Perrozi, B. et al. "DeepWalk: Online Learning of Social Representations." SIGMOD, 2014.

³Wu, Z. et al. "A Comprehensive Survey on Graph Neural Networks." IEEE Trans. Neural Networks and Learning Systems, 2021.

EXAMPLE: FRIEND RECOMMENDATIONS & NETWORK GROWTH



- **Scenario: Social Network “People You May Know”**
 - **Applications:** SNs use LP to suggest new connections.
 - **GA:** Algorithms rank potential friends.
 - **SNA interpretation:** Principles like Triadic Closure and Homophily explain success.
- **Key insights:**
 - **Increased connectivity:** PYMK boosts new friendship formation.
 - *Impact:* A large fraction of new connections originates from suggestions.
 - **Eerily accurate:** Often surfaces people users genuinely know but hadn't searched for.
 - *Validation:* Confirms algorithms effectively capture real-world social ties.
 - **Actionable growth:** Drives network density and user engagement.
 - **Beyond social:** Professional networking (e.g., LinkedIn) and fraud detection (hidden links).

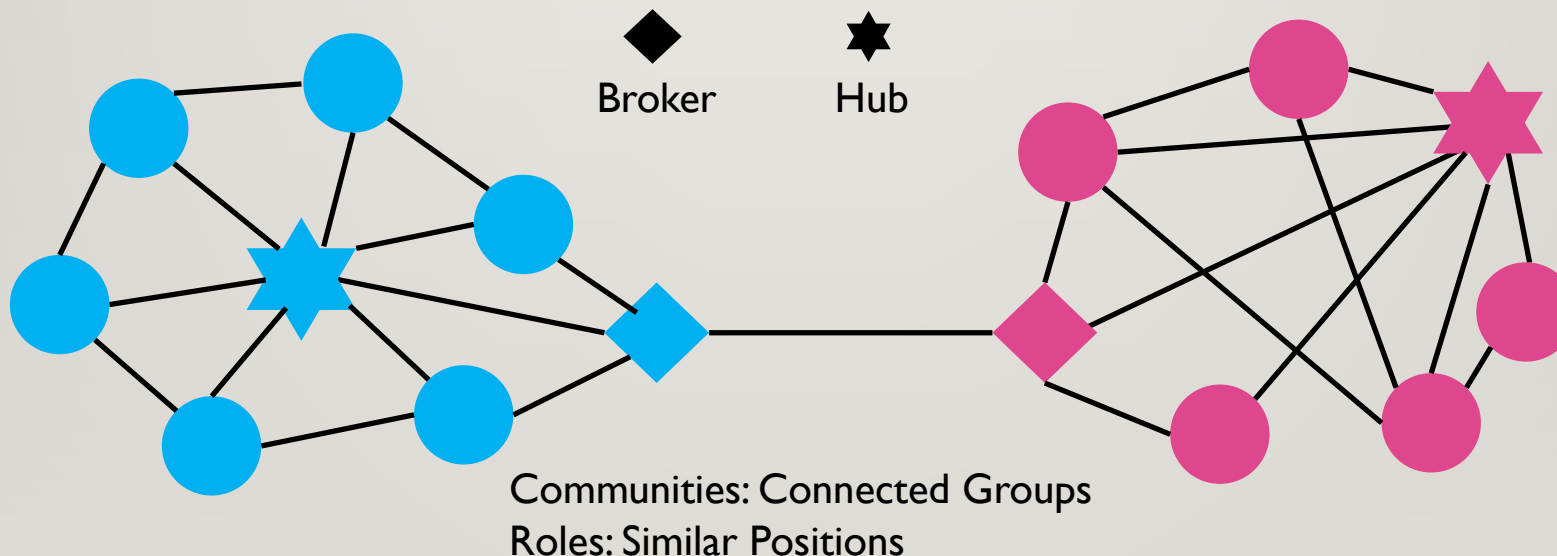
ROLES IN NETWORKS – BEYOND COMMUNITIES

- **What are roles?**

- Position in network structure characterized by a particular pattern of ties.
- Nodes in the same role may not be connected to each other.
- Often transcend communities.

- **Why identify roles?**

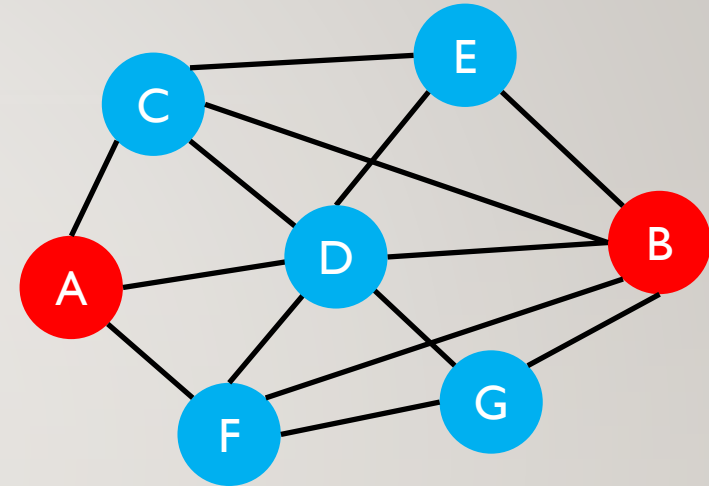
- Simplification
- Transfer of Insights
- Detection of Anomalies
- Social insight



SNA ON ROLES: STRUCTURAL EQUIVALENCE & BLOCKMODELS

- **Structural Equivalence (SE):**

- **Definition:** Two nodes have exactly the same neighbours.
- **Interpretation:** Nodes are substitutable in terms of network position.
- *Strict condition, rarely met in large networks.*



Structural Equivalence

SNA ON ROLES: STRUCTURAL EQUIVALENCE & BLOCKMODELS

- **Structural Equivalence (SE):**

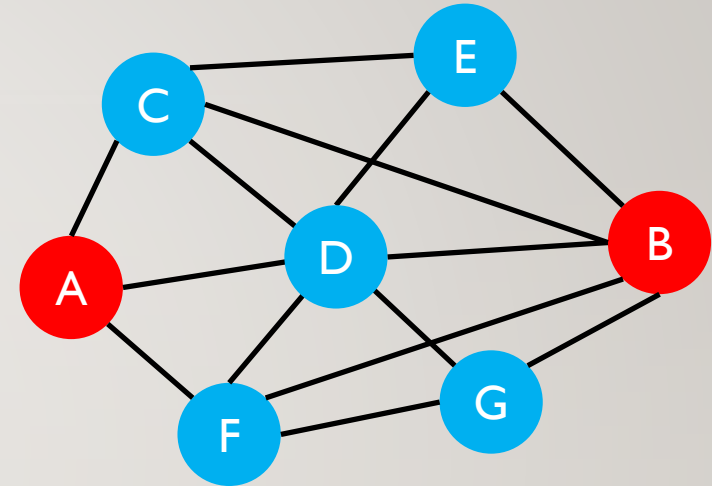
- **Definition:** Two nodes have exactly the same neighbours.
- **Interpretation:** Nodes are substitutable in terms of network position.
- *Strict condition, rarely met in large networks.*

- **Regular Equivalence (RE):**

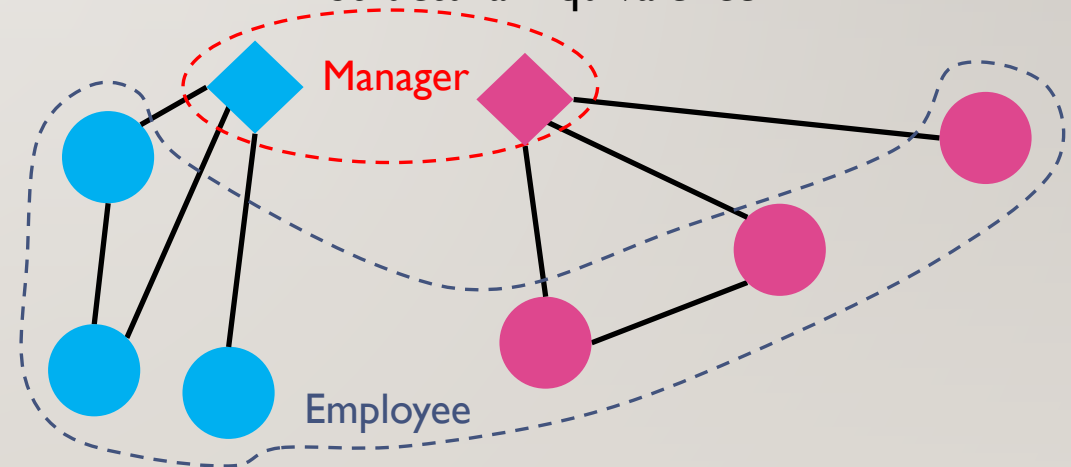
- **Definition:** Nodes connect to others in the *same roles*.
- **Interpretation:** Captures the essence of a “role”.

- **Blockmodeling:**

- **Technique:** Cluster nodes such that patterns of ties *between* clusters are uniform.
- *Goal:* Reduce network complexity while preserving structural patterns.



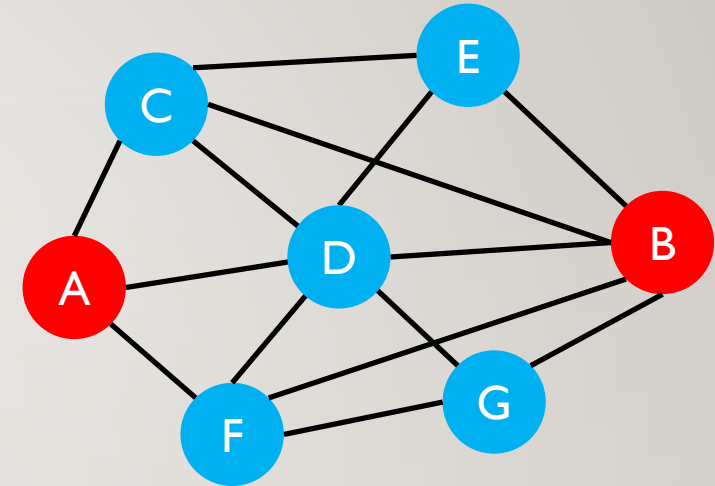
Structural Equivalence



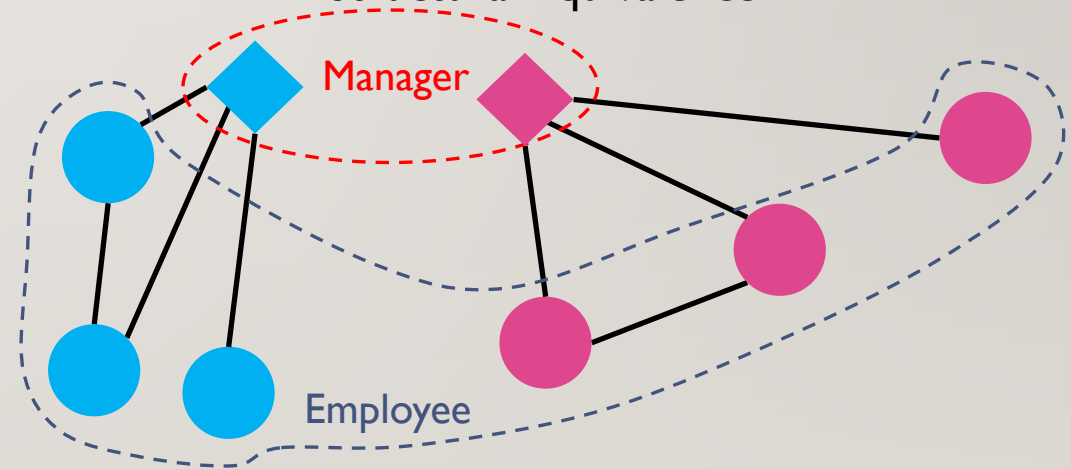
Regular Equivalence/Blockmodeling

SNA ON ROLES: STRUCTURAL EQUIVALENCE & BLOCKMODELS

- **Structural Equivalence (SE):**
 - **Definition:** Two nodes have exactly the same neighbours.
 - **Interpretation:** Nodes are substitutable in terms of network position.
 - *Strict condition, rarely met in large networks.*
- **Regular Equivalence (RE):**
 - **Definition:** Nodes connect to others in the *same roles*.
 - **Interpretation:** Captures the essence of a “role”.
- **Blockmodeling:**
 - **Technique:** Cluster nodes such that patterns of ties between clusters are uniform.
 - *Goal:* Reduce network complexity while preserving structural patterns.
- **SNA Interpretation:** Give meaning to roles by examining node composition and inter-role connections.



Structural Equivalence

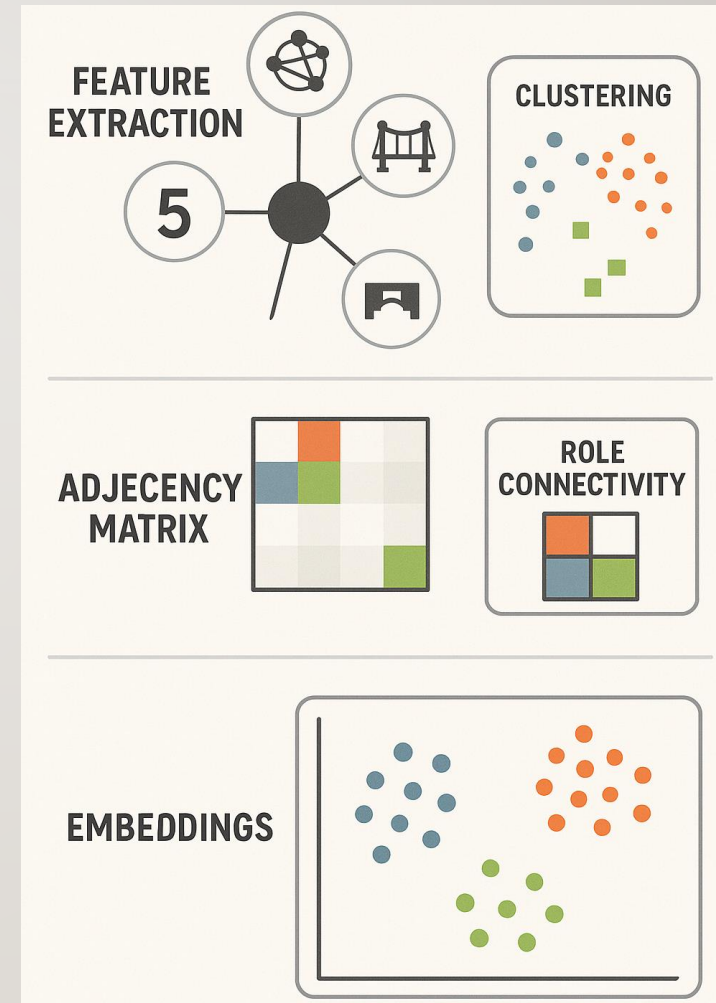


Regular Equivalence/Blockmodeling

GA FOR ROLE DISCOVERY

- **Feature extraction + clustering:**

- **Process:** Compute a set of structural features for each node (e.g., degree, ego network density, motif counts, centrality measures).
- **Clustering:** Group nodes with similar feature vectors into roles.
- **Advantage:** Flexible, captures various structural patterns, scalable.



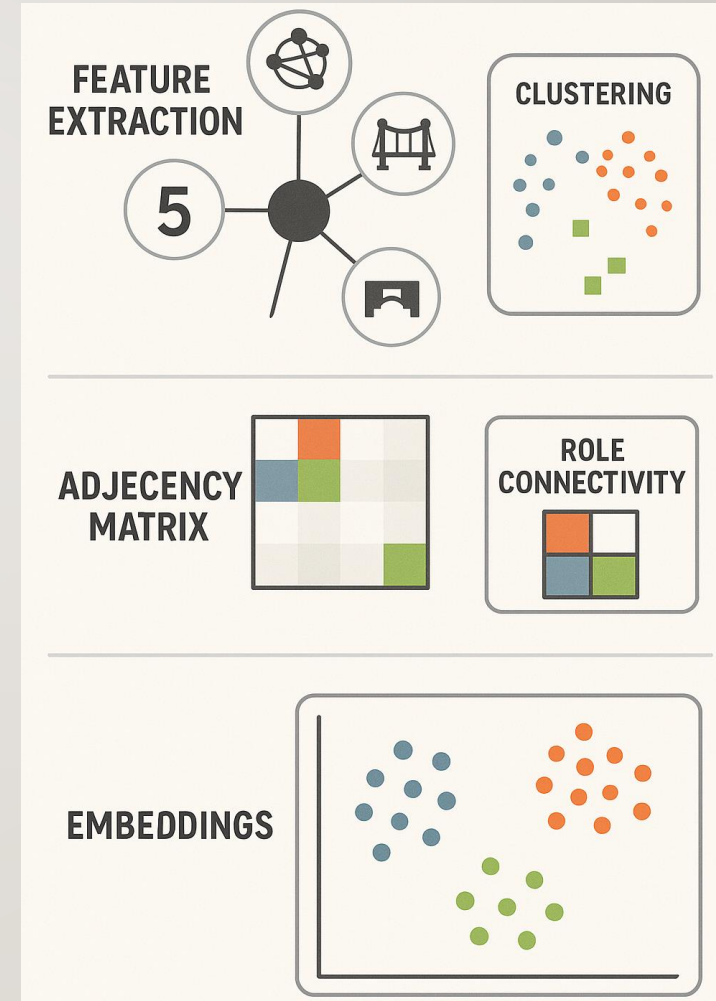
GA FOR ROLE DISCOVERY

- **Feature extraction + clustering:**

- **Process:** Compute a set of structural features for each node (e.g., degree, ego network density, motif counts, centrality measures).
- **Clustering:** Group nodes with similar feature vectors into roles.
- **Advantage:** Flexible, captures various structural patterns, scalable.

- **Matrix factorization of adjacency:**

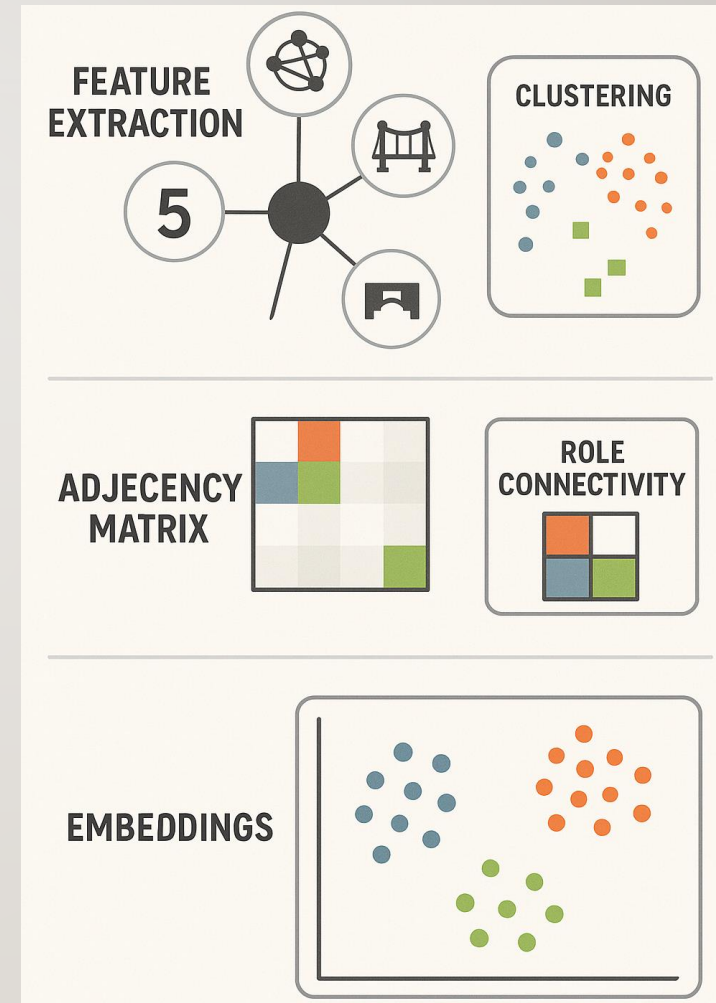
- Model node connections as a function of their underlying roles.
 - Factorize the network's adjacency matrix to uncover latent role patterns.



Henderson, K. et al. "RoIX: Structural Role Extraction & Mining in Large Graphs." *KDD*, 2012.
Mitra, A. et al. "RoIX: A Unified Non-Negative Matrix Factorization Framework for Semi Supervised Learning on Graphs." *SDM*, 2020.

GA FOR ROLE DISCOVERY

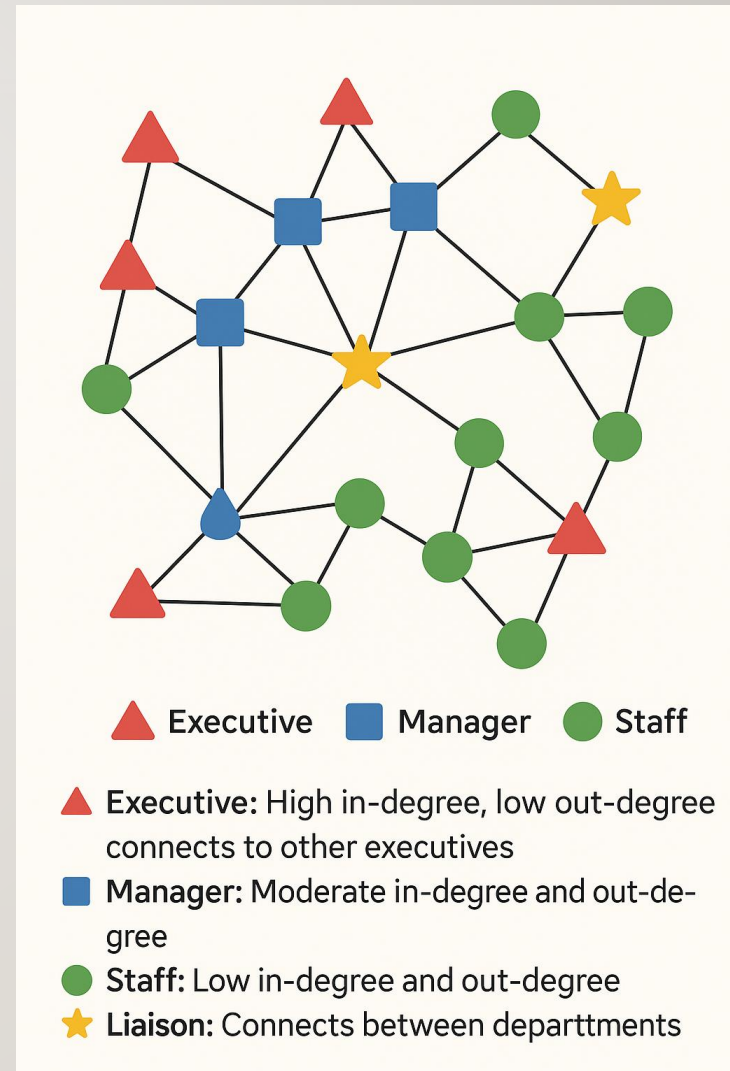
- **Feature extraction + clustering:**
 - **Process:** Compute a set of structural features for each node (e.g., degree, ego network density, motif counts, centrality measures).
 - **Clustering:** Group nodes with similar feature vectors into roles.
 - **Advantage:** Flexible, captures various structural patterns, scalable.
- **Matrix factorization of adjacency:**
 - Model node connections as a function of their underlying roles.
 - Factorize the network's adjacency matrix to uncover latent role patterns.
- **Graph embeddings for roles:**
 - Generate embeddings that capture structural similarity.
 - Cluster embeddings to identify role groups.
 - **Advantage:** Can find roles even in disconnected components.



Henderson, K. et al. "RoIX: Structural Role Extraction & Mining in Large Graphs." *KDD*, 2012.
Mitra, A. et al. "RoIX: A Unified Non-Negative Matrix Factorization Framework for Semi Supervised Learning on Graphs." *SDM*, 2020.
N. K. Ahmed et al., "Role-Based Graph Embeddings." *IEEE TKDE*, 2022.

EXAMPLE: DISCOVERING ROLES IN AN ORGANIZATION

- **Scenario: Enron email network analysis**
 - **Network:** Email communication among Enron employees.
 - **GA:** Role-mining algorithm applied.
 - **SNA Interpretation:** Mapped detected roles to organizational functions and behaviours.
- **Key insights:**
 - **Formal roles confirmed:** Identified roles largely aligned with formal hierarchy.
 - *Validation:* Confirmed algorithm's ability to capture known structure.
 - **Emergent “liaison” role:** Discovered a distinct role of individuals with cross-departmental connections.
 - *Insight:* These were often executive assistants or IT staff – unofficial integrators.
 - **Actionable simplification:** Reduced complexity of 150+ nodes to 4-5 key roles.
 - *Utility:* Simplified analysis, highlighted critical communication paths, revealed potential vulnerabilities.



SUMMARY: SNA – GA PERSPECTIVES ON THE FIVE TASKS

Task	SNA Focus	GA Focus
Influence	Interpret via centrality, actor position	Influence maximization, scalable centrality
Communities	Cohesive subgroups, attributes	Modularity, Louvain, scalable clustering
Diffusion	Threshold models, weak ties theory	Epidemic simulations, super-spreader ID
Link prediction	Homophily, triadic closure	Similarity heuristics, ML embeddings
Roles	Structural/regular equivalence	Feature extraction, embeddings, clustering

AGENDA

Introduction & Foundations

Five Examples

Graph Analytics Platforms

- Programming models
- Computation models

Integration, Tools & Methodologies

Challenges & Future Directions

DEALING WITH SCALE

Almost all systems are scale-out

DEALING WITH SCALE

Almost all systems are scale-out

Can data parallel platforms (e.g., MapReduce) be used for graph analytics?

DEALING WITH SCALE

Almost all systems are scale-out

Can data parallel platforms (e.g., MapReduce) be used for graph analytics?

Yes, but not a good idea

- These systems do not perform well on iterative tasks

DEALING WITH SCALE

Almost all systems are scale-out

Can data parallel platforms (e.g., MapReduce) be used for graph analytics?

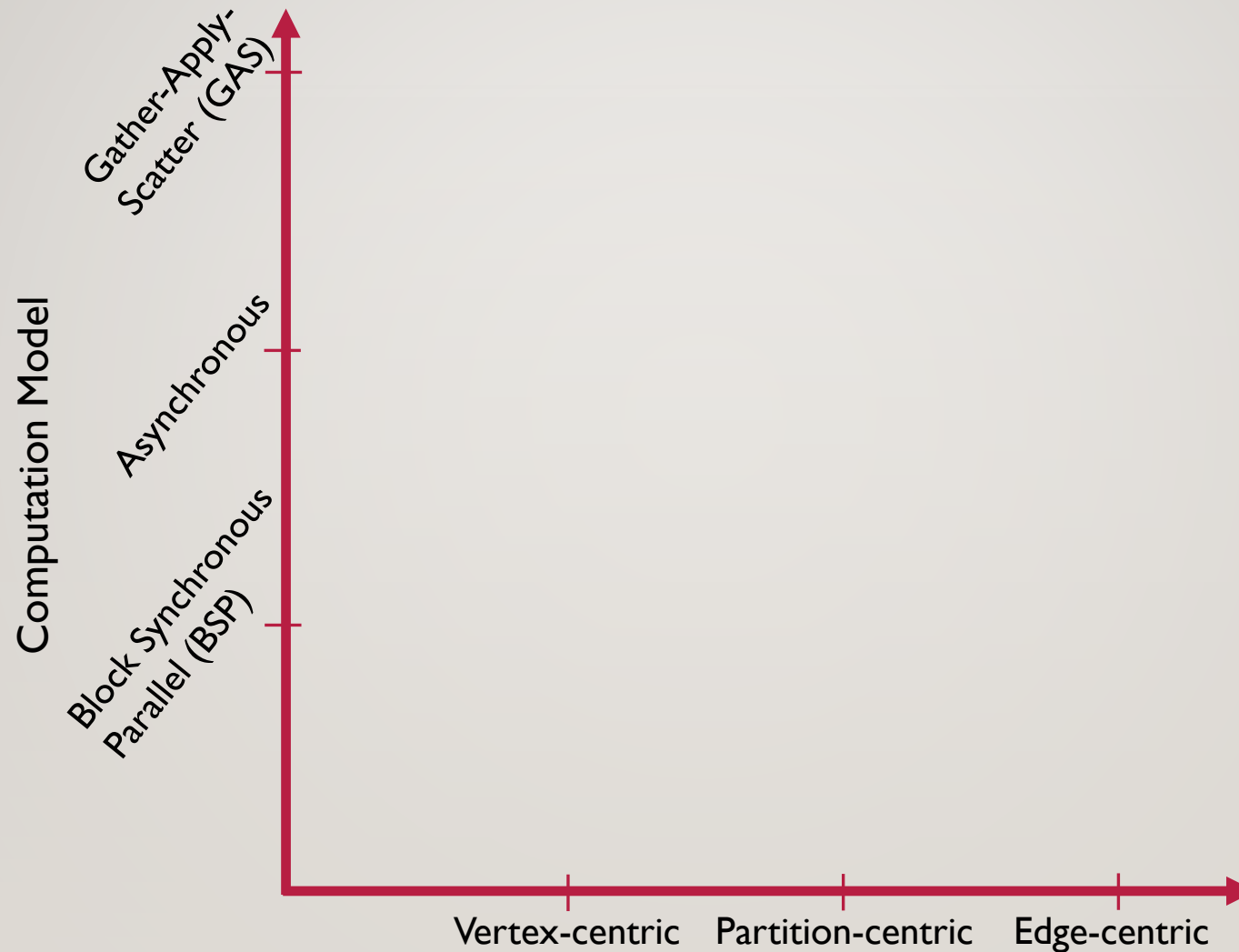
Yes, but not a good idea

- These systems do not perform well on iterative tasks

Specialized platforms

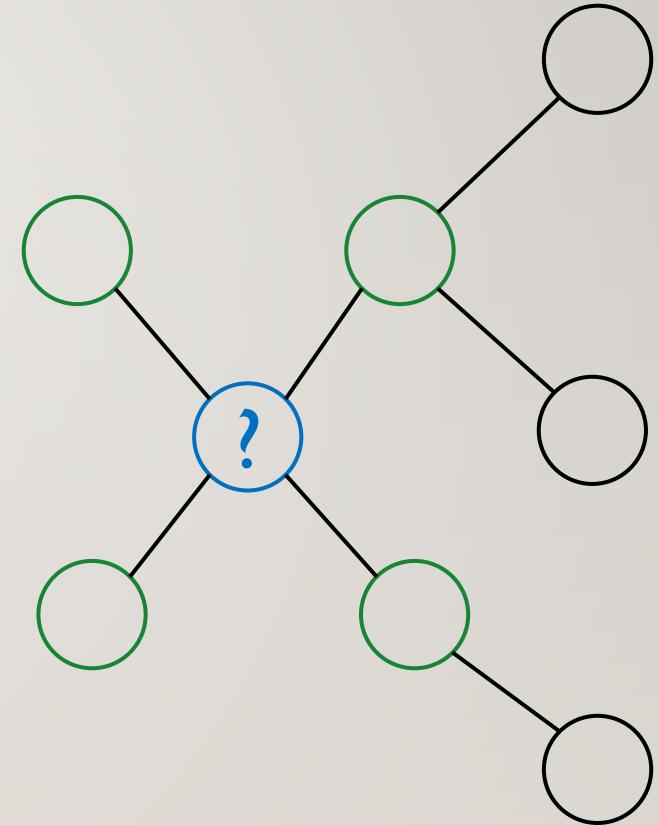
- Some ideas incorporated into graph DBMSs

CLASSIFICATION OF GRAPH ANALYTICS SYSTEMS



PROGRAMMING MODELS

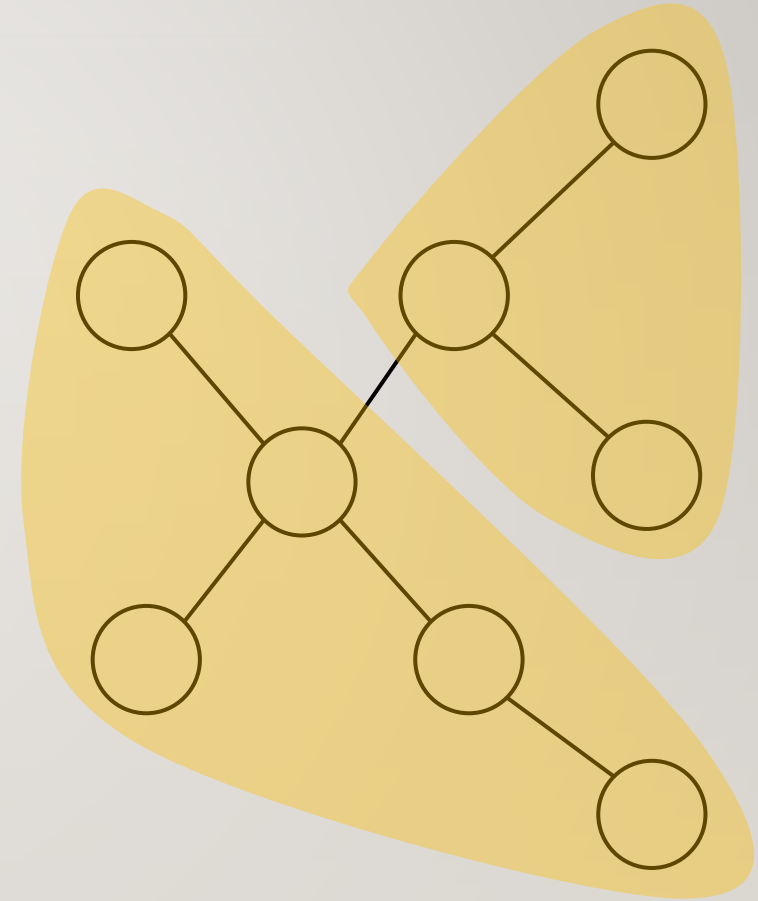
- Vertex-centric¹
 - Computation on a vertex is the focus.
 - “Think like a vertex”.
 - Vertex computation depends on its own state + states of its neighbours.
 - Functions:
 - `Compute(vertex v)`
 - `GetValue()`, `WriteValue()`



¹Malewicz, G. et al. “Pregel: a system for large-scale graph processing.” *SIGMOD*, 2010.

PROGRAMMING MODELS

- Vertex-centric¹
- Partition-centric (Block-centric)²
 - Computation on an entire partition is specified.
 - “Think like a block” or “Think like a graph”.
 - Aim is to reduce the communication cost among vertices.

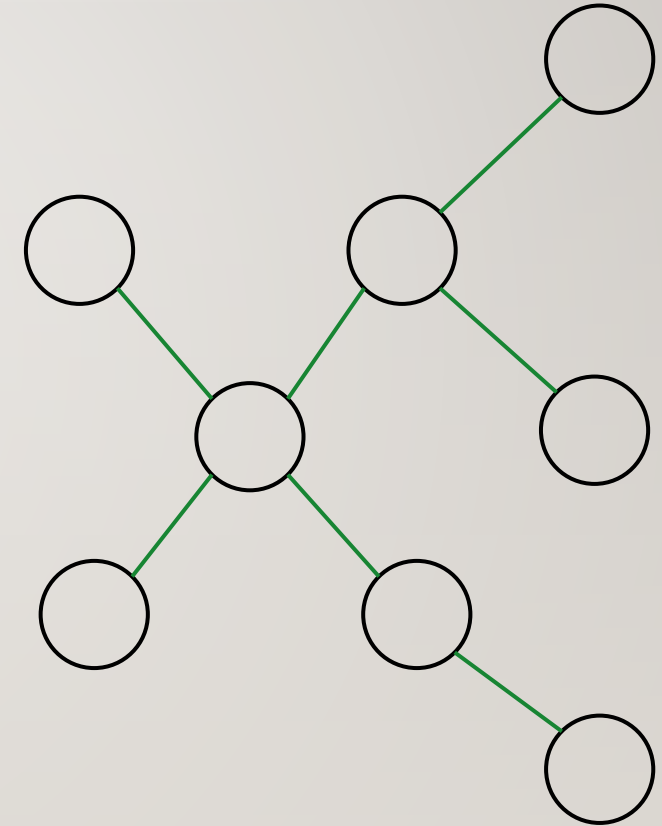


¹Malewicz, G. et al. “Pregel: a system for large-scale graph processing.” *SIGMOD*, 2010.

²Yan, D. et al. “ Blogel: A block-centric framework for distributed computation on real-world graphs.” *PVLDB*, 2014.

PROGRAMMING MODELS

- Vertex-centric¹
- Partition-centric (Block-centric)²
- Edge-centric³
 - Computation is specified on each edge rather than on each vertex or block
 - `Compute(edge e)`



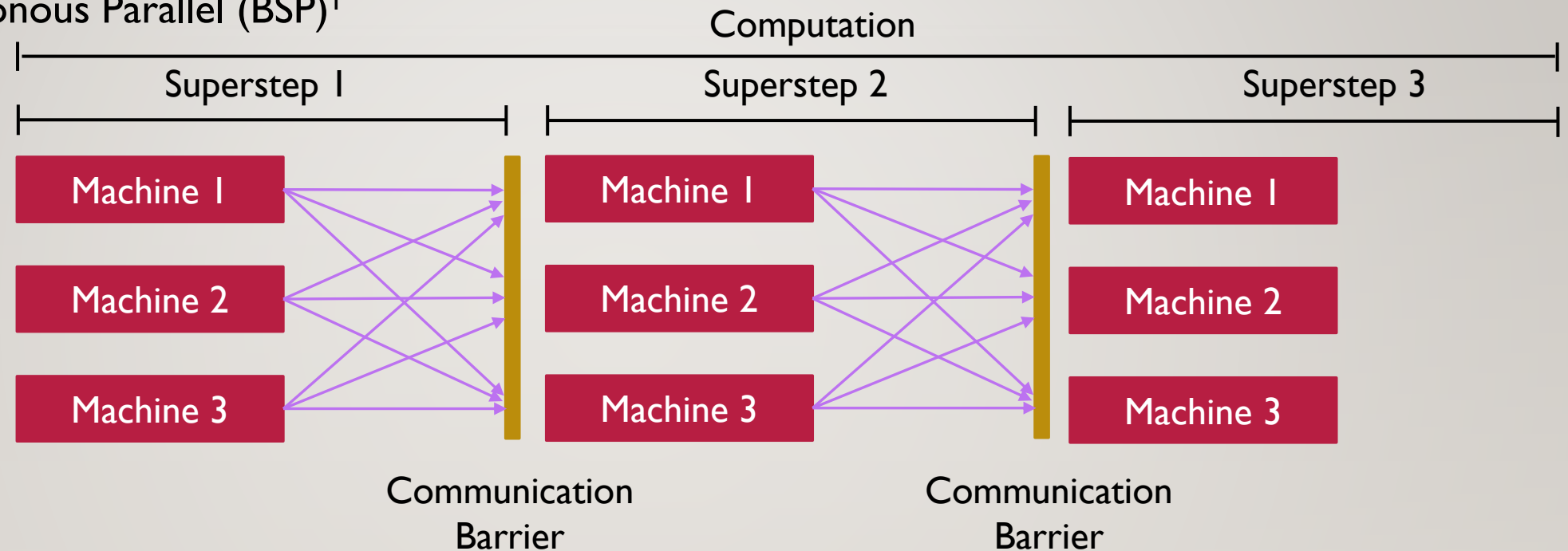
¹Malewicz, G. et al. "Pregel: a system for large-scale graph processing." *SIGMOD*, 2010.

²Yan, D. et al. "Blogel: A block-centric framework for distributed computation on real-world graphs." *PVLDB*, 2014.

³Roy, A. et al. "X-stream: edge-centric graph processing using streaming partitions." *SOSP*, 2013.

COMPUTATION MODELS

- Bulk Synchronous Parallel (BSP)¹



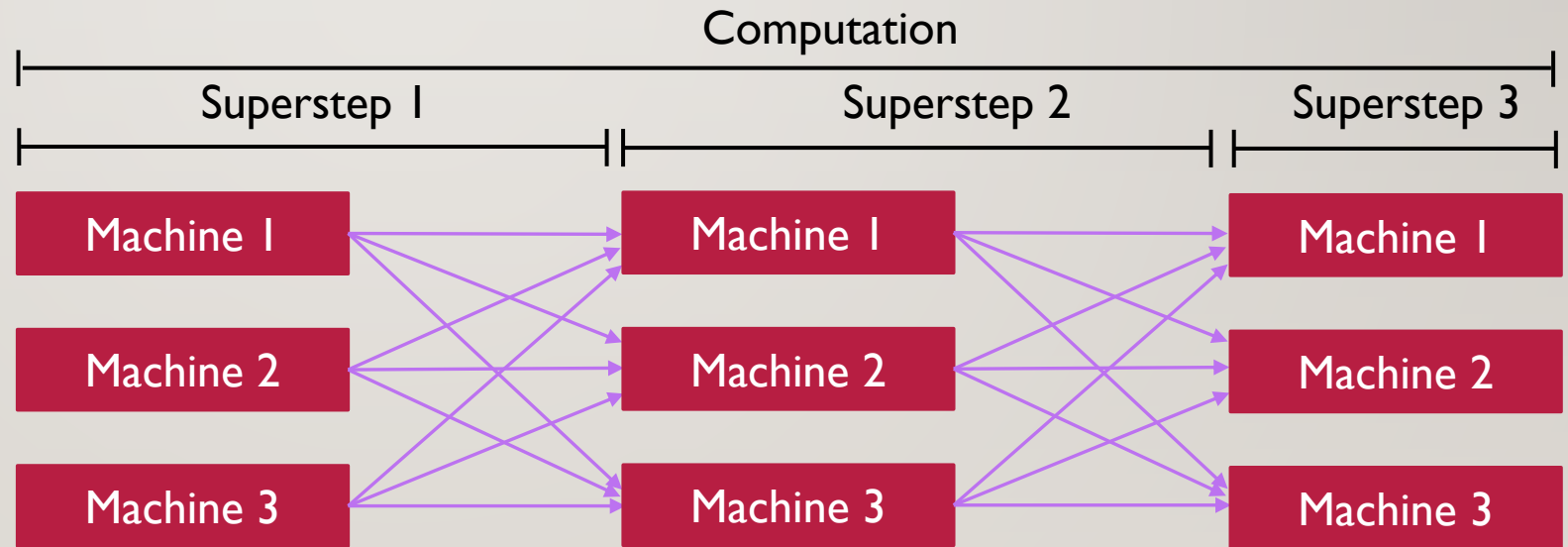
Each machine performs
computation
on its graph partition

At the end of each superstep
results are **pushed** to other
workers

¹Valiant, L.G. "A bridging model for parallel computation." CACM, 1990.

COMPUTATION MODELS

- Bulk Synchronous Parallel (BSP)¹
- Asynchronous Parallel²
 - No communication barriers.
 - Uses the most recent state values.
 - Implemented using distributed locking.



¹Valiant, L.G. "A bridging model for parallel computation." *CACM*, 1990.

²Han, M. and Daudjee, K. Giraph unchained: Barrierless asynchronous parallel execution in Pregel-like graph processing systems." *PVLDB*, 2015.

COMPUTATION MODELS

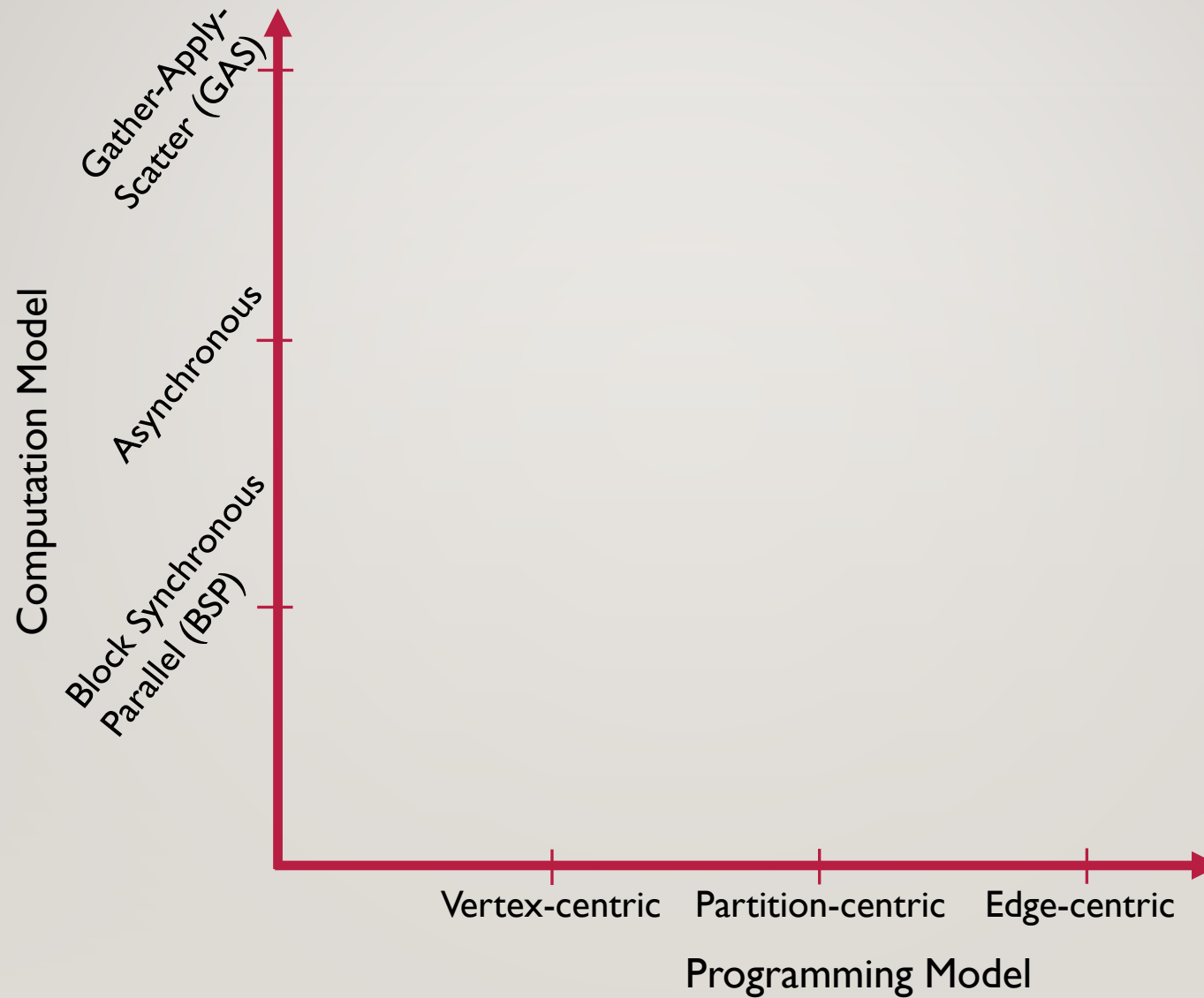
- Bulk Synchronous Parallel (BSP)¹
- Asynchronous Parallel²
- Gather-Apply-Scatter (GAS)³
 - Similar to BSP, but pull-based
 - Gather: pull state
 - Apply: Compute function
 - Scatter: Update state
 - Updates of states separated from scheduling

¹Valiant, L.G. “A bridging model for parallel computation.” *CACM*, 1990.

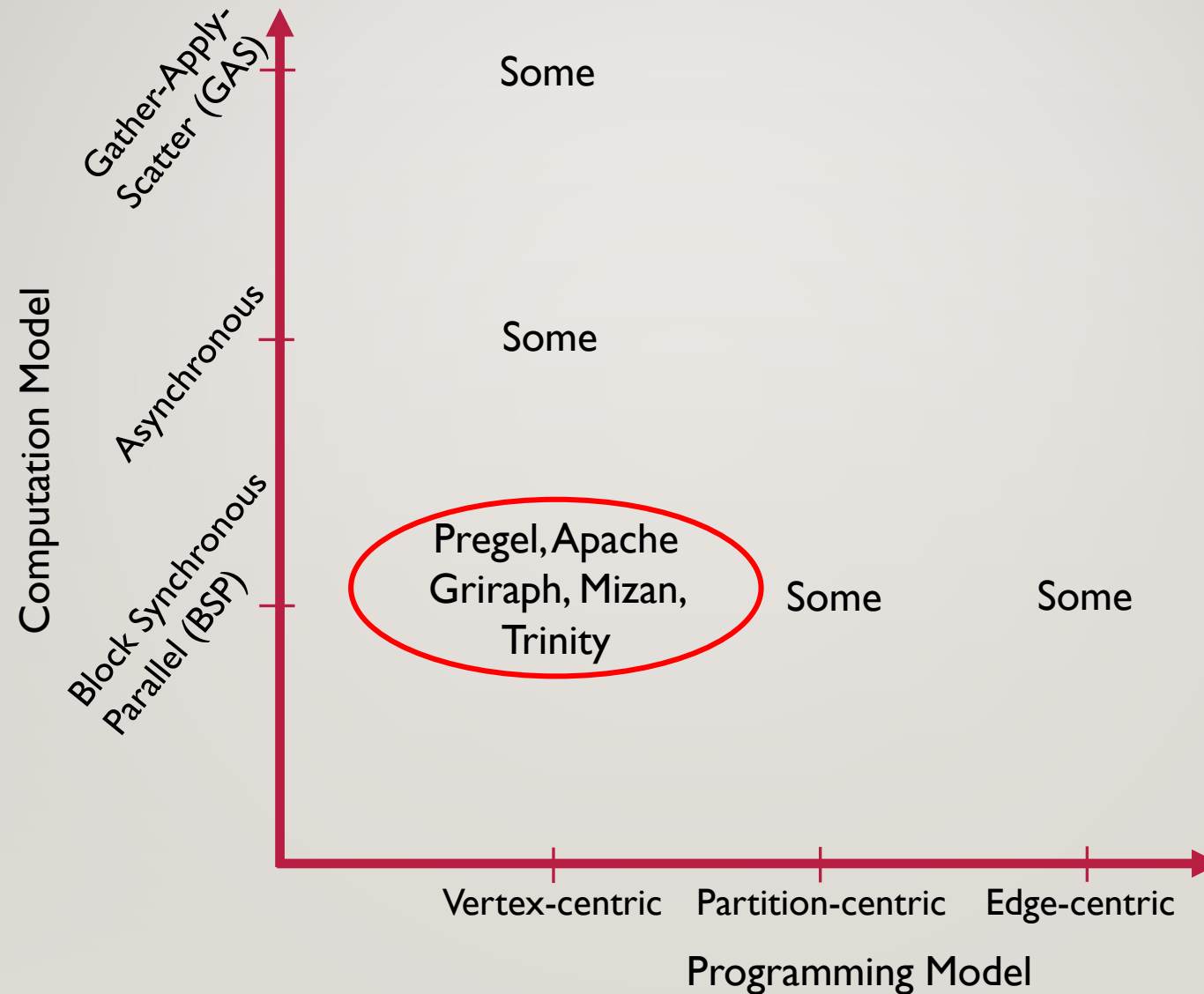
²Han, M. and Daudjee, K. Giraph unchained: Barrierless asynchronous parallel execution in Pregel-like graph processing systems.” *PVLDB*, 2015.

³Low, Y., et al. “GraphLab: new framework for parallel machine learning.” *Conf. on Uncertainty in AI*, 2010.

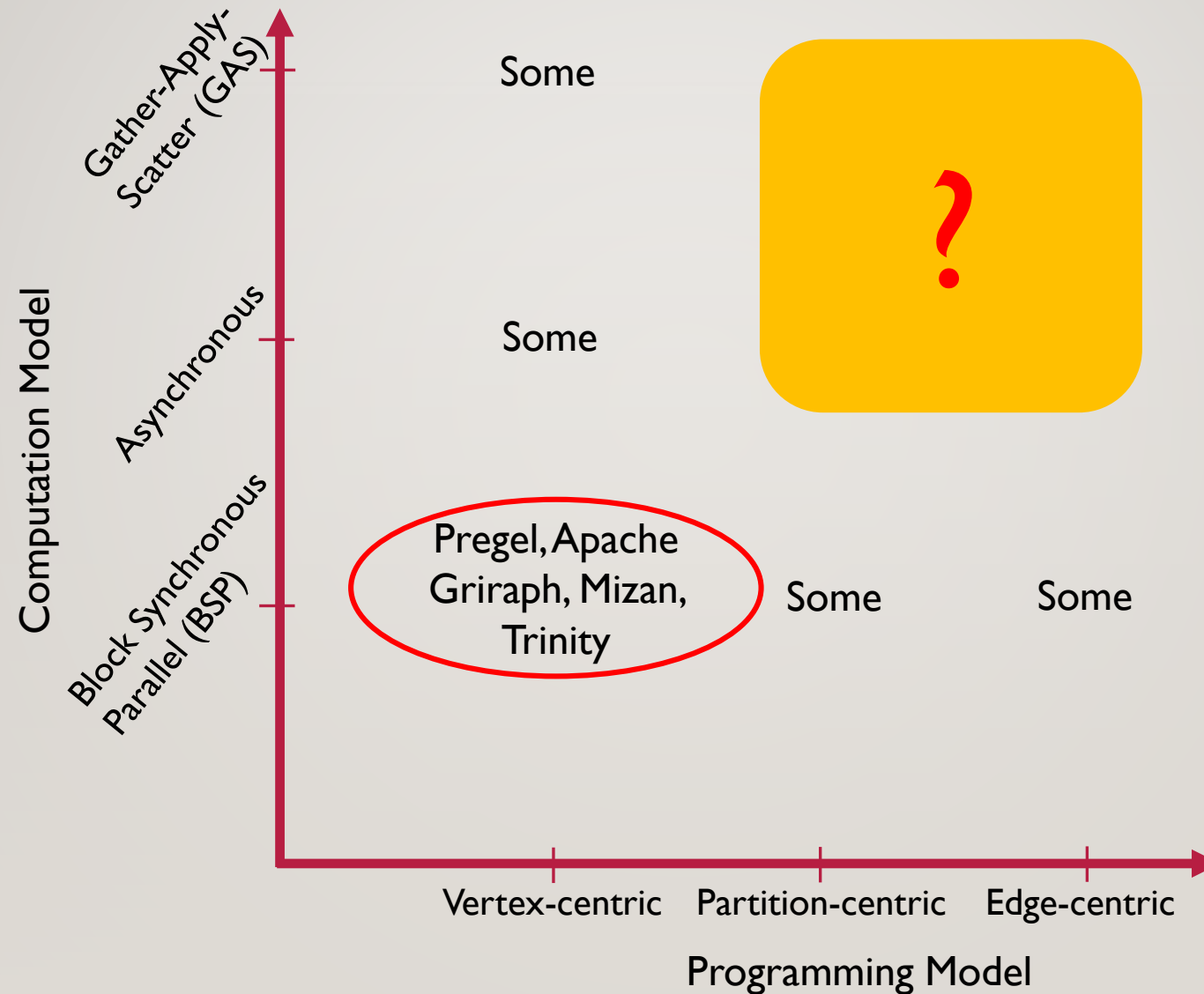
CLASSIFICATION OF GRAPH ANALYTICS SYSTEMS



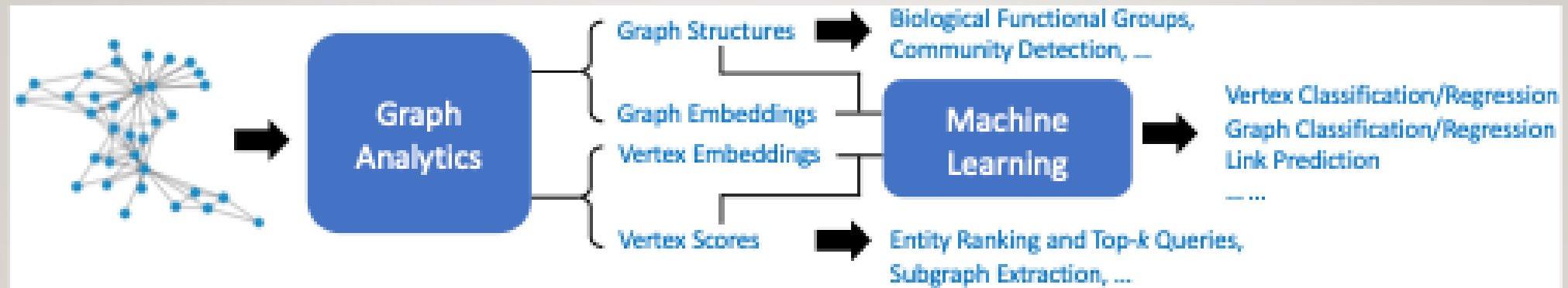
CLASSIFICATION OF GRAPH ANALYTICS SYSTEMS



CLASSIFICATION OF GRAPH ANALYTICS SYSTEMS



GRAPH ANALYTICS - MACHINE LEARNING PIPELINE



	Problem Type		Search Approach		
	(Pattern-to-Instance)	(Instance-to-Pattern)			
	Subgraph Finding	FSM (A Big Graph)	FSM (transaction database)	Subgraph Materialization	Sugraph Backtracking
					GPU Support
Arabesque	✓	✓		✓	
RStream	✓	✓		✓	
Pangolin	✓	✓		✓	EG
Peregrine	✓	✓		✓	
Fractal	✓	✓			✓
AutoMine	✓	✓		✓	
G-thinker	✓			✓	
G-Miner	✓			✓	
GraphPi	SE/SM			✓	
GraphZero	SE/SM			✓	
T-DFS	SE/SM			✓	EG
GSI, cuTS, STMatch, EGSM	SE/SM			✓	EG
PBE, VSGM, SGSI	SE/SM			✓	PT
G ² -AIMD	✓			✓	PT
DistGraph		✓		✓	
ScaleMine		✓			✓
T-FSM		✓		✓	
PrefixFPM			✓	✓	

* SE/SM = Subgraph Enumeration/Matching

* EG = Entire Graph (Loaded to GPU)

* PT = A Partition of a Graph (Loaded to GPU) Each Time

Systems	Optimizations				
	Operators Graph Data Communication	Model Scheduling and Synchronization	Other Optimizations	Full-Graph GNN	
Euler		✓			
AliGraph		✓			
DistDGL	✓				
AGL	✓				
P ³	✓	✓	✓		
NeutronStar		✓			
ByteGNN	✓	✓			
DGCL	✓			✓	
BGL	✓	✓			
Sancus		✓	✓		
Dorylus		✓	✓	✓	
DistGNN	✓				✓
HongTu	✓				✓

AGENDA

Introduction & Foundations

Five Examples

Graph Analytics Platforms

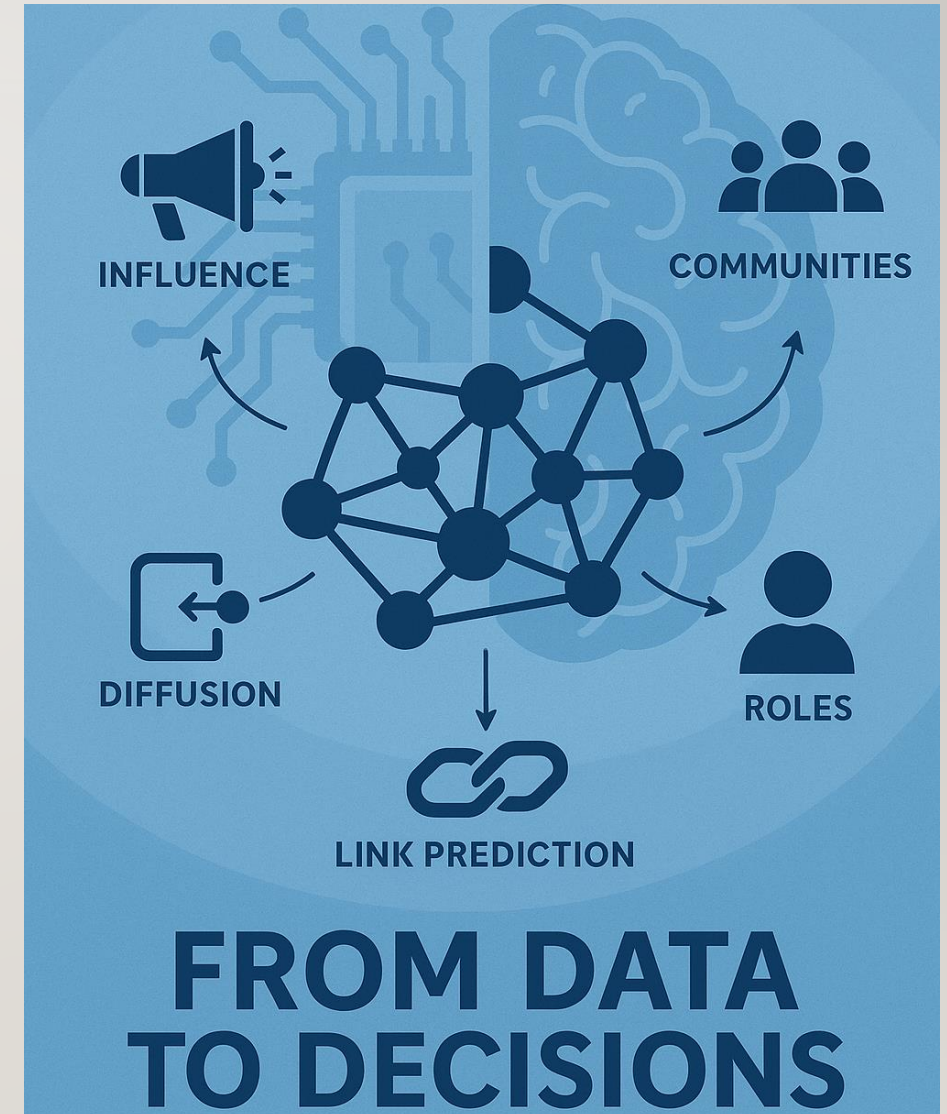
Integration, Tools & Methodologies

- Synthesizing insights
- Enabling technologies

Challenges & Future Directions

INTEGRATING PERSPECTIVES – THE FULL NETWORK PICTURE

- **Holistic understanding:**
 - **Influence:** Pinpoint key individuals at scale, understand *why* they matter.
 - **Communities:** Detect meaningful groups, interpret their social significance.
 - **Diffusion:** Simulate spread, strategize.
 - **Link prediction:** Anticipate network evolution, generate accurate recommendations.
 - **Roles:** Abstract networks into functional positions, simplify complexity.
- **Cross-topic connections:**
 - Influence & Communities: Identify influencers *within* each group.
 - Diffusion & Communities: Understand how group boundaries affect spread.
 - Link Prediction & Roles: Predict ties based on structural similarity.
- **Unified data & platforms:** Facilitates seamless analysis.
- **Outcome:** Transform complex networks into *actionable knowledge and smarter decisions*.



TOOLS & PLATFORMS – FROM DATA TO INSIGHT

- **Graph DBMSs & GA Engines**

- *Benefit:* Scalability to billions of edges, unified data system.

- **Network analysis libraries**

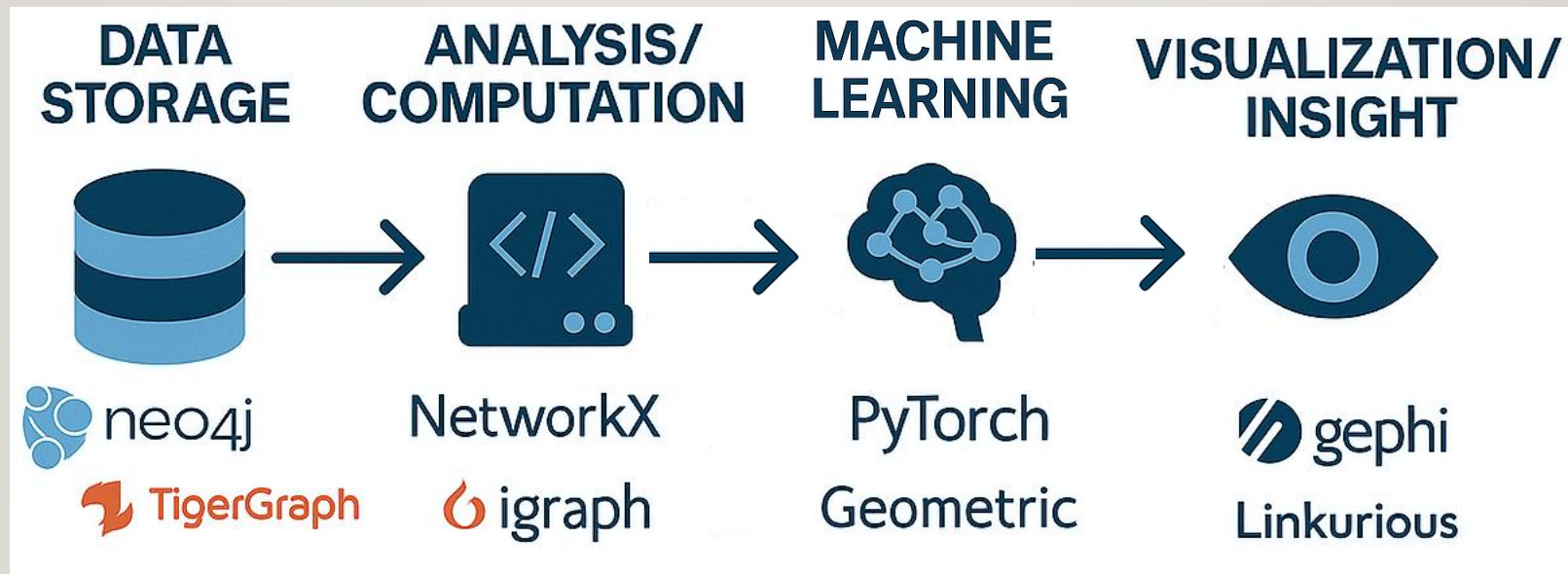
- **NetworkX (Python), igraph (R/Python), Graph-tool:** For smaller to medium networks, rapid prototyping, SNA measures.
- *Benefit:* Flexible, integrate with general data science workflows.

- **Graph Machine Learning frameworks**

- **PyTorch Geometric, DGL (Deep Graph Library)**
Build and train GNNs & advanced graph ML models.
- *Benefit:* Leverage GPU acceleration, cutting-edge prediction and representation learning.

- **Visualization & Interactive tools**

- **Gephi, Cytoscape, Linkurious:** Visual exploration, interactive analysis, computation with human insight.



AGENDA

Introduction & Foundations

Five Examples

Graph Analytics Platforms

Integration, Tools & Methodologies

Challenges & Future Directions

- Key challenges
- Future outlook

CHALLENGES: DATA QUALITY, PRIVACY & SCALE

- **Data quality & completeness:**

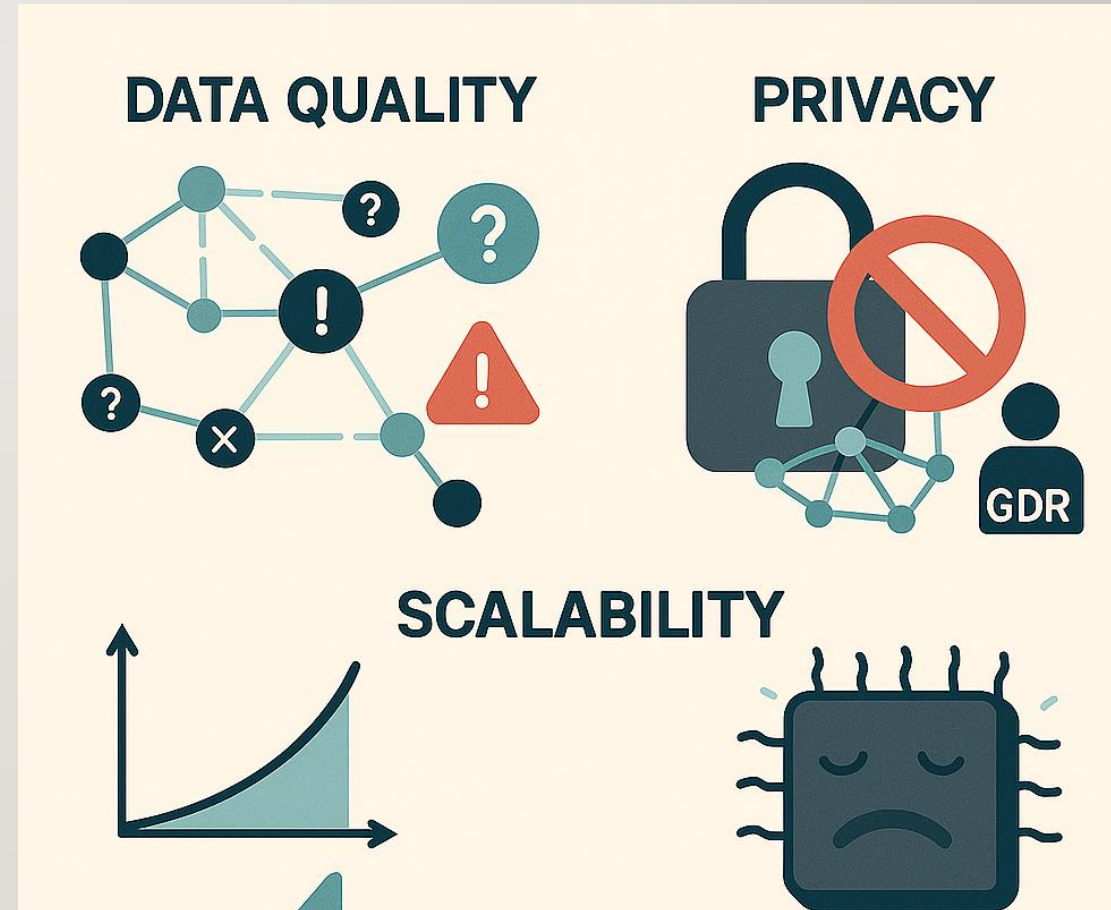
- Incomplete or noisy real-world network data.
- Can bias centrality, community structure, link prediction.

- **Privacy & ethics:**

- Social network data is sensitive; powerful inferences can reveal private information.
- Ethical concerns: manipulative use of influence, potential for surveillance.
- Regulatory restrictions (e.g., GDPR & PIPEDA) and platform API limitations.

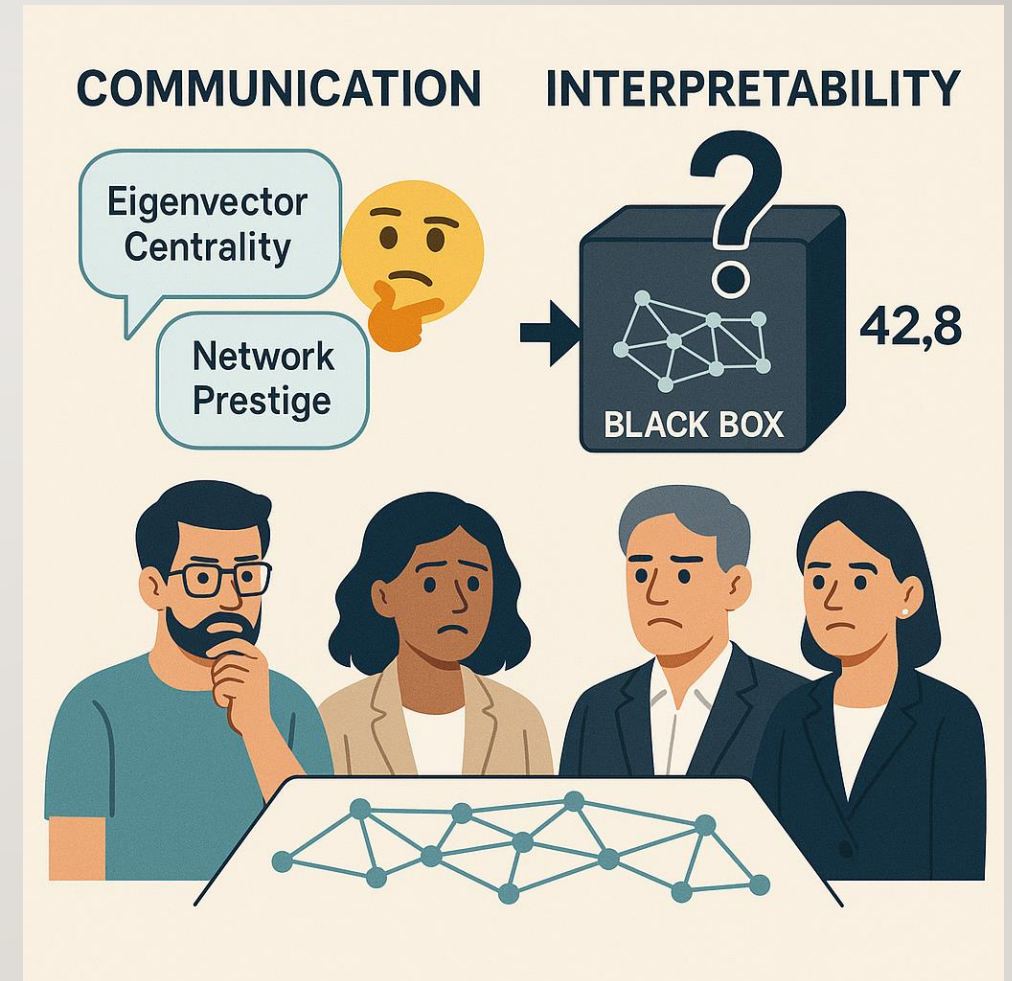
- **Scalability & complexity:**

- Truly huge networks still pose challenges.
- Some algorithms don't scale linearly (e.g., exact betweenness).
- Dynamic/real-time analysis challenging.
- Multi-relational/multi-layer networks: current tools often handle one relation type.



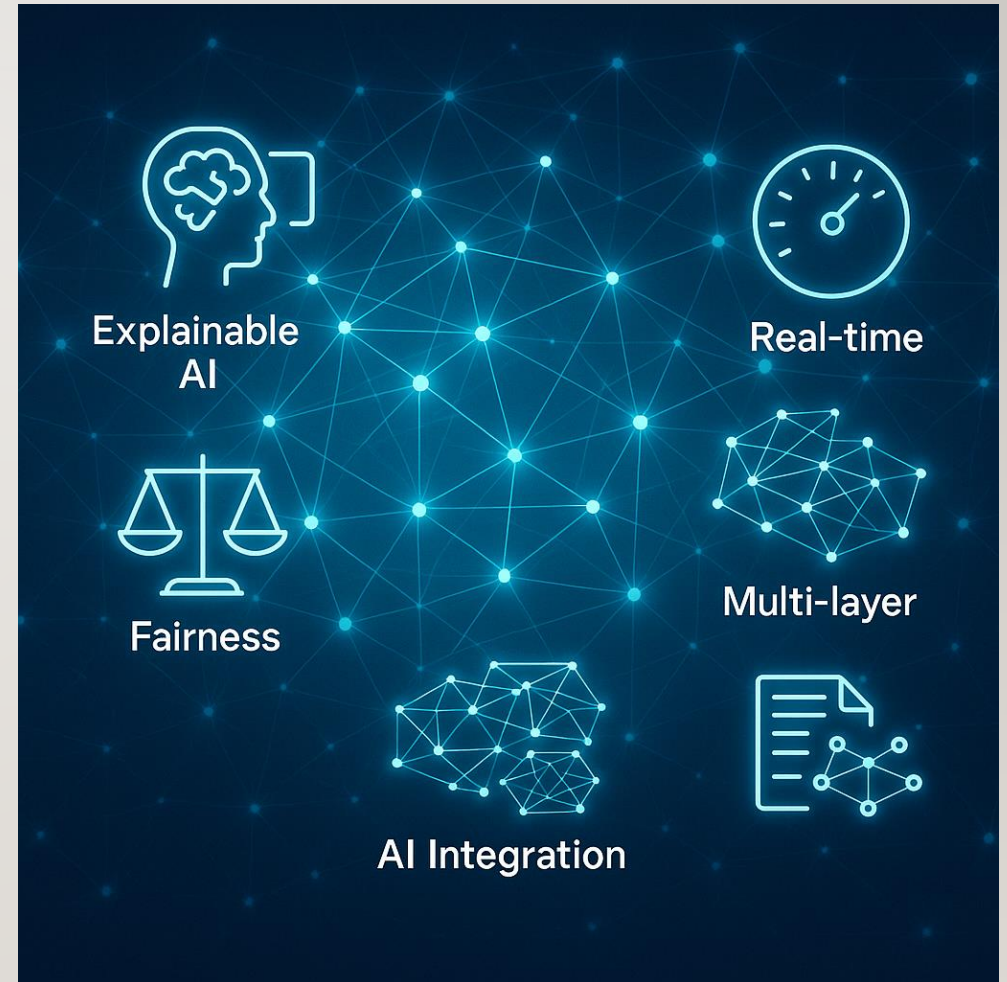
CHALLENGES: INTERDISCIPLINARY GAPS & INTERPRETABILITY

- **Interdisciplinary communication**
 - **Language gap:** Different jargon.
 - **Priority differences:** Interpretability (SNA) vs. Accuracy/Efficiency (GA).
 - **Organizational silos:** Data science teams vs. social science researchers.
- **Interpretability of advanced models**
 - Black-box models
 - Need for explanation
 - Bias reinforcement
- **Convincing stakeholders**
 - Network insights can be counter-intuitive.
 - Organizational inertia or politics.
 - Ethical considerations in practical application.



FUTURE DIRECTIONS: SMARTER, FAIRER NETWORK ANALYSIS

- **Explainable & transparent Graph ML:**
 - Developing GNNs that provide *reasons* for predictions.
- **Bias & fairness mitigation:**
 - Incorporating fairness constraints into algorithms.
- **Dynamic & real-time analysis:**
 - Algorithms handling streaming graph updates..
- **Multi-layer network analysis:**
 - Better handling of networks with multiple relationship types.
- **Deeper integration with other AI:**
 - Combining network analysis with NLP, Computer Vision for richer context.
- **Ethical frameworks & cross-discipline training:**
 - Developing guidelines for responsible use.



ACKNOWLEDGEMENTS



ACKNOWLEDGEMENTS



Gen AI tools have been used in structuring this presentation and in the creation of some of the infographics



CONCLUSION – TOWARD A DEEPER NETWORK UNDERSTANDING

- **Restated thesis:** The combination of SNA and graph analytics provides a powerful lens for understanding and shaping networked systems.
- **Key outcomes recap:**
 - Pinpoint *influencers* and strategize with them.
 - Uncover *community structure* and understand silos.
 - Simulate *diffusion* and plan interventions.
 - Predict *new connections* and foster growth.
 - Abstract networks into *roles* for simplified understanding.
- **Implication:** Invaluable for organizations and society in an interconnected world.
- **Collaboration:** Requires diverse expertise – computer scientists, sociologists, domain experts.
- **Final vision:** Transform big data into deep insights, leading to smarter decisions and stronger communities → data science.

