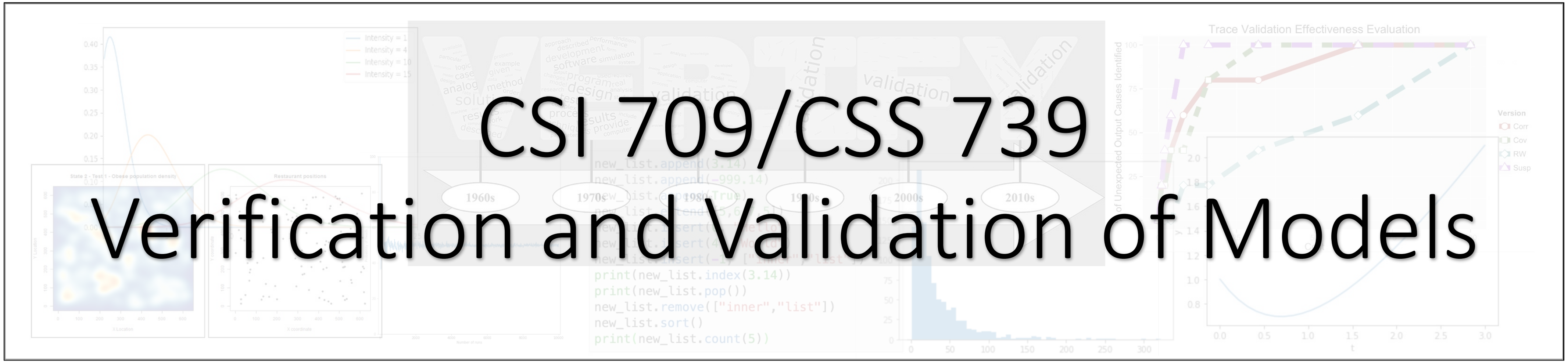


# CSI 709/CSS 739
# Verification and Validation of Models

## Validation of Network Models

Dr. Hamdi Kavak
Computational and Data Sciences Department

http://www.hamdikavak.com
hkavak@gmu.edu

George Mason University

Center for Social Complexity

# Weekly schedule

- Week 13 (today Nov 13): Validation of Network Models
- Week 14 (Nov 20): Ethics of Model Design and Use

------------------------

- Week 15 (Nov 27): Project presentations
- Week 16 (Dec 4): READING DAYS – no lecture
- Week 17: No meeting — final paper due on Dec 8 @ 11:59 pm

# Final paper presentations

- Slides are **due at noon** on the day of your presentation (Nov 27)
- Presentation length: 25 minutes (main) + 5 minutes (Q&A)

# Short Writing Assignments (SWAs)

- SWA 9: Was due today
- SWA 10: Due next week (Nov 20)
- If you are missing any previous assignments, you can finish them until next class (Nov 20).
  - I will be done with all SWA grading next week

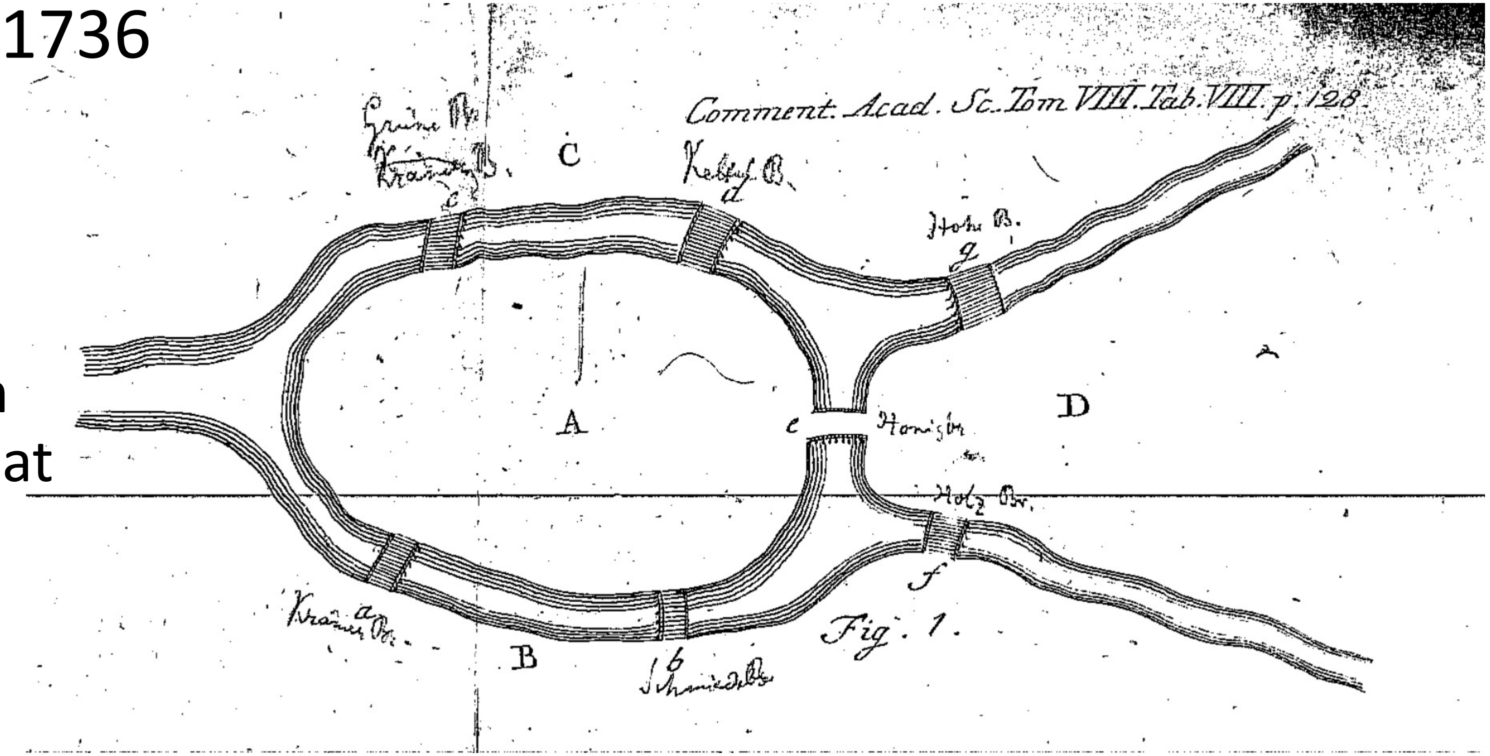
# Validation of Network Models

# Network models

- A formal way of representing and studying complex systems based on the concept of **graphs**.

# Network models

- A formal way of representing and studying complex systems based on the concept of **graphs**.
- The origin of graphs dates to 1736
- Leonhard Euler proposed the problem called "The Seven Bridges of Königsberg"
  - Can seven bridges be visited in one pass with the condition that we finish at the start point?




Source: Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 128-140.

# Building blocks

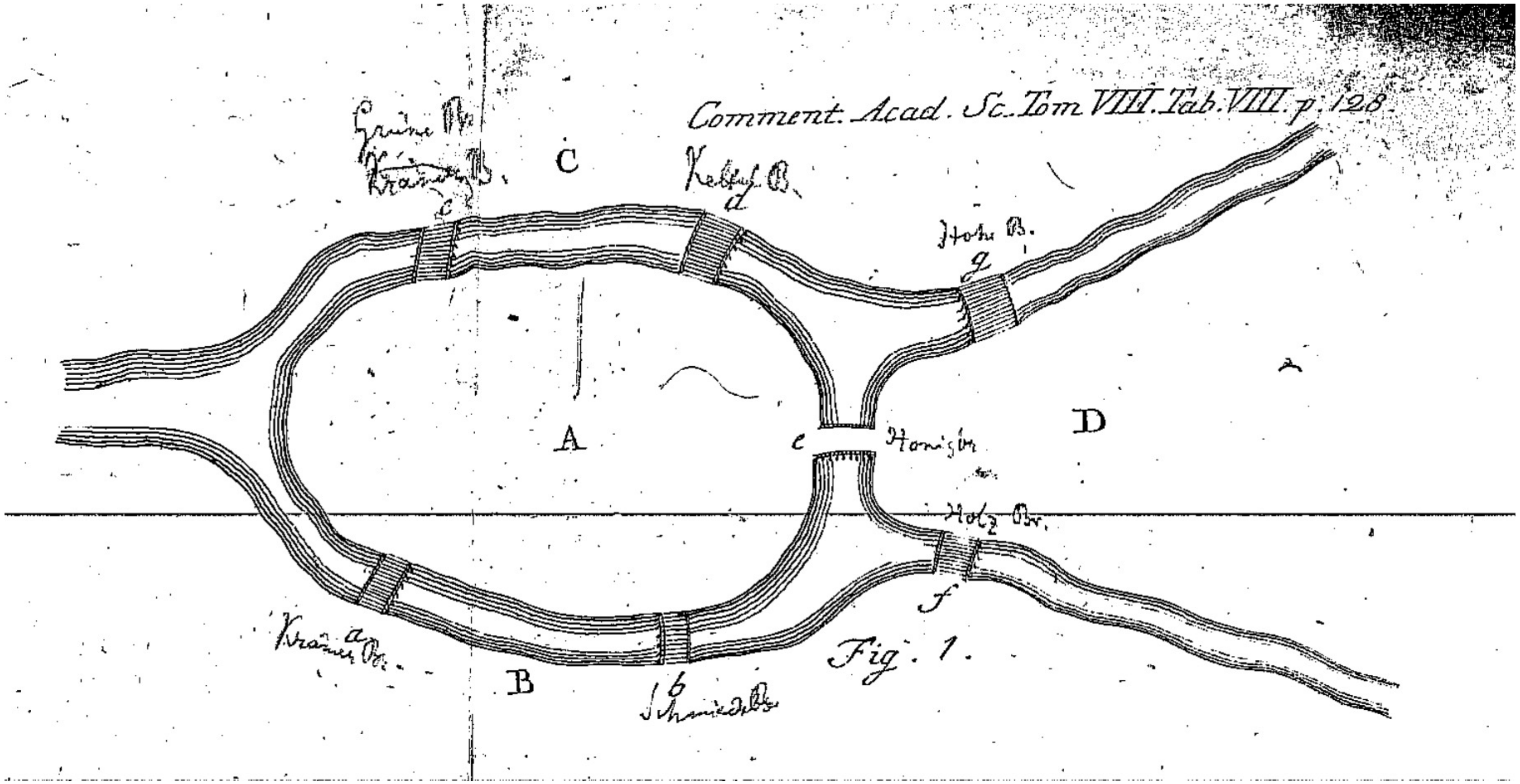
**Nodes (vertices)**

**Links (edges)**

# Building blocks

**Nodes (vertices)**

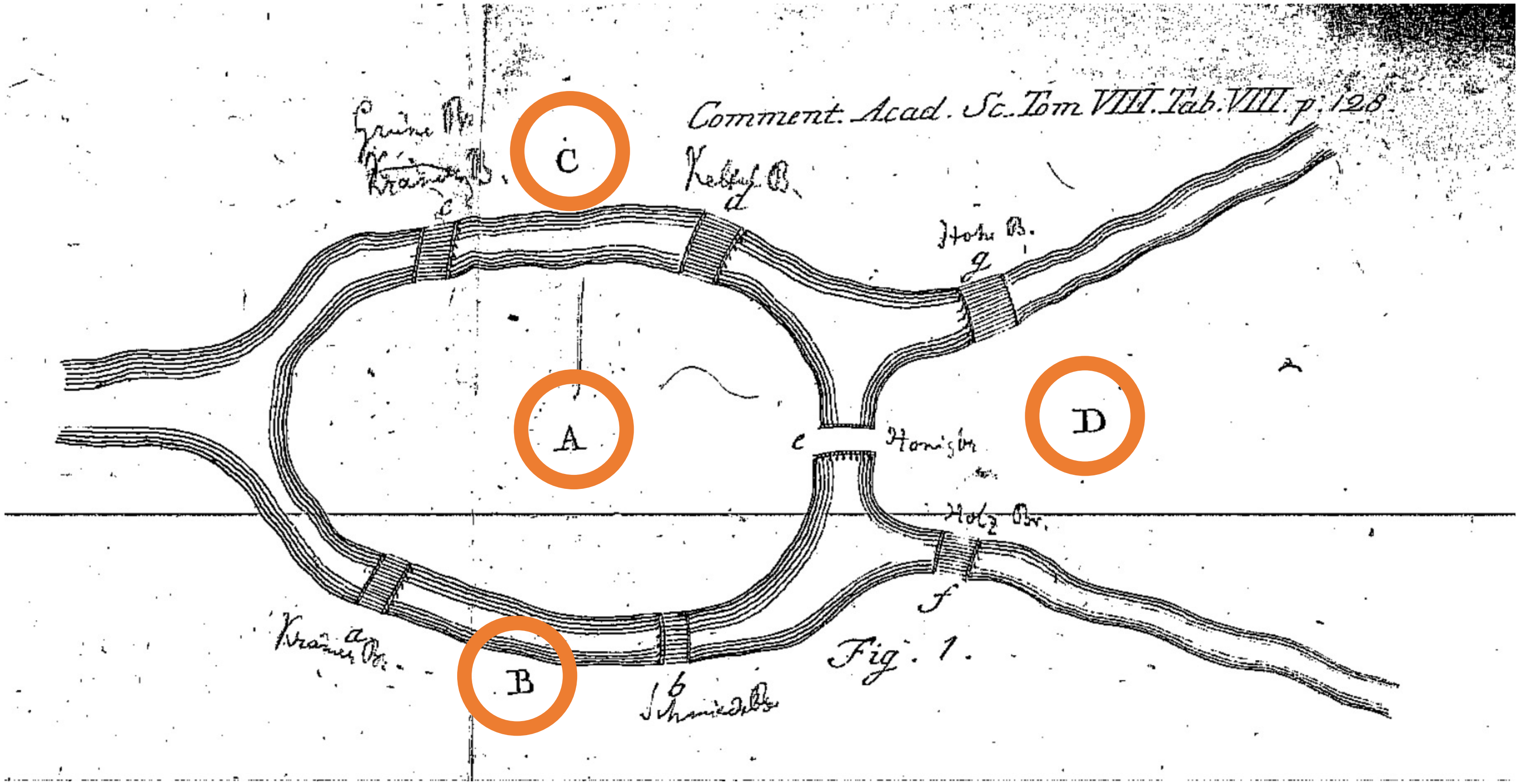
**Links (edges)**




Source: Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 128-140.

# Building blocks

Nodes (vertices) — Links (edges)

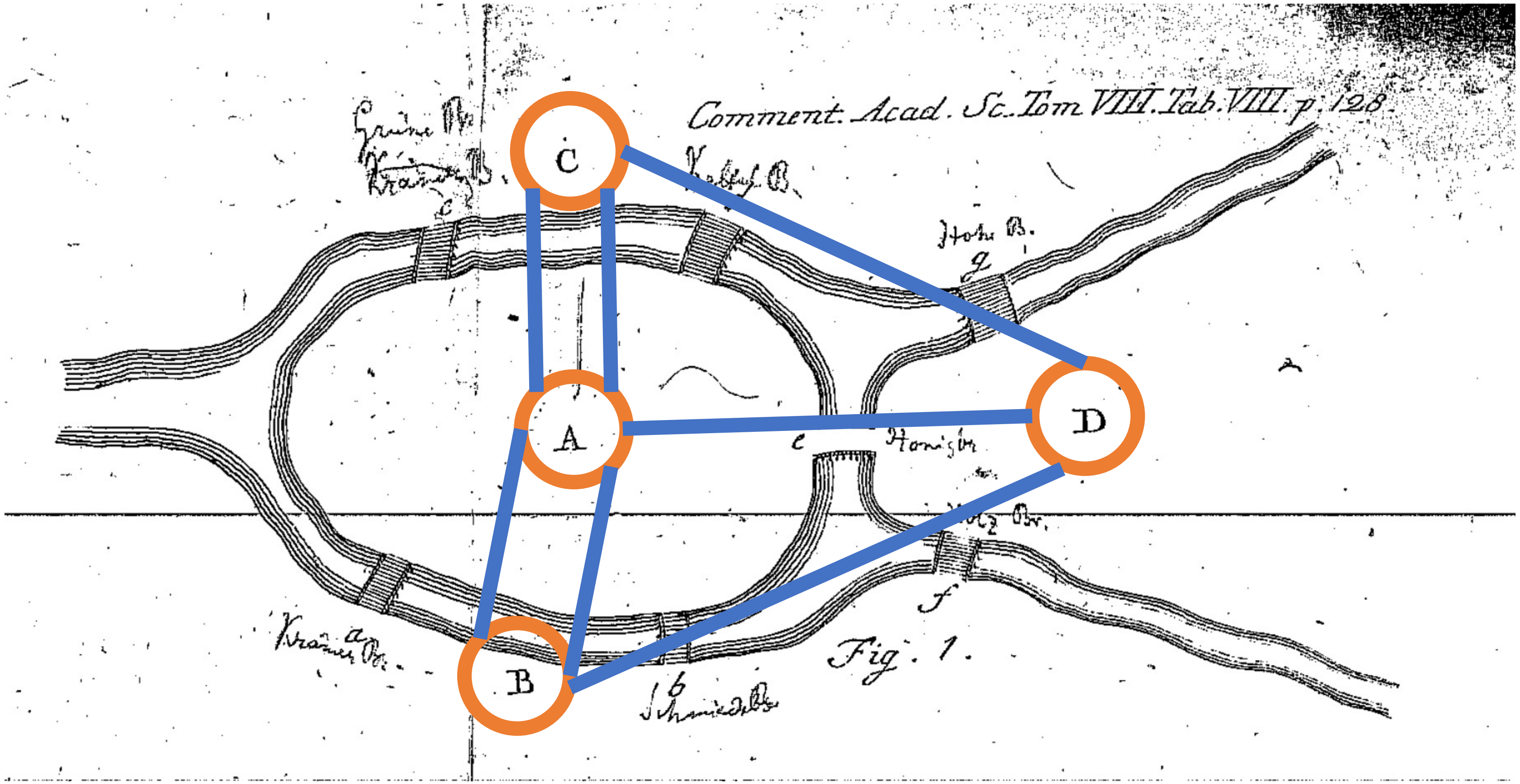



Source: Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 128-140.

# Building blocks

Nodes (vertices)

Links (edges)



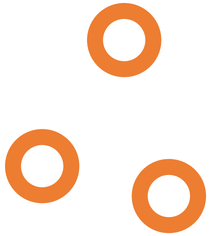

Source: Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 128-140.

# Building blocks

- **Nodes** can be anything that can **connect** to other nodes
  - E.g.: people, countries, cities, bus stations, buildings, molecules, web pages, social media users, cell towers, positions in an organization, military units.

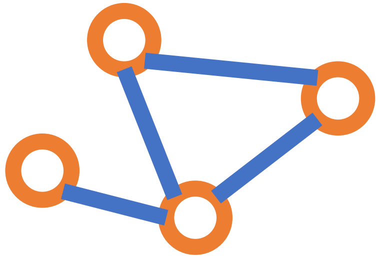
# Building blocks

- **Nodes** can be anything that can **connect** to other nodes
  - E.g.: people, countries, cities, bus stations, buildings, molecules, web pages, social media users, cell towers, positions in an organization, military units.
- Some graph types



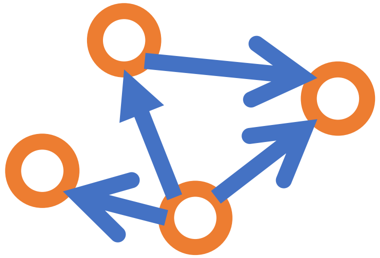
**Null graph**

no links



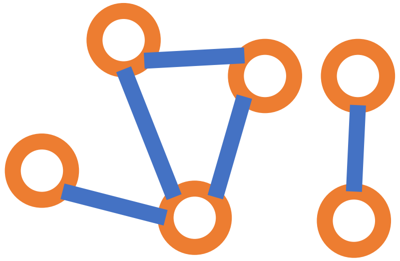
**Undirected graph**

No directions in links



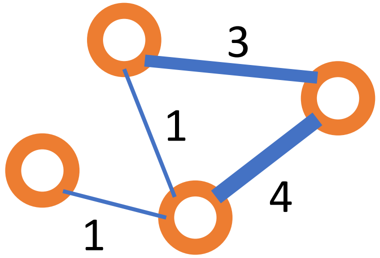
**Directed graph**

Directions in links



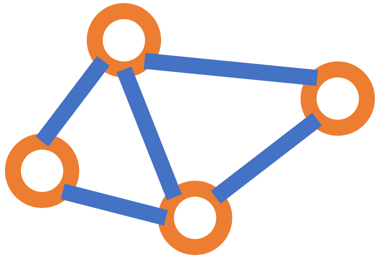
**Disconnected graph**

Cannot visit all nodes in one pass




**Weighted graph**

Links are not equal

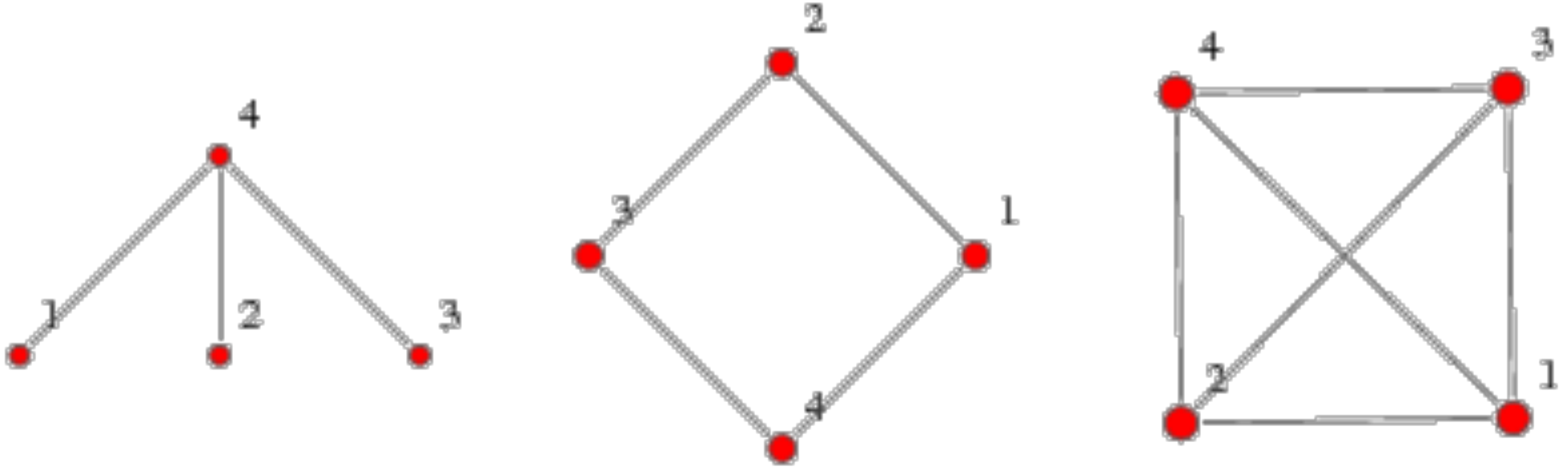


**Cyclic graph**

At least one cycle exists

# Adjacency matrix

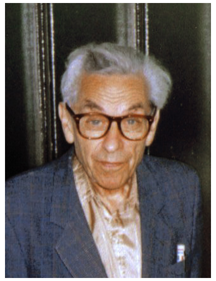
- A matrix to represent the connectivity between nodes
- Denoted by $A$




$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

Source: https://mathworld.wolfram.com/AdjacencyMatrix.html

# Some influential people in network science*



Pául Erdős

Alfréd Rényi

Alessandro Vespignani

Albert-László Barabási

Steven Strogatz

Réka Albert

Duncan J. Watts

Aaron Clauset

Petter Holme

*20th and 21st century

# Small world networks

Source: https://www.youtube.com/watch?v=NRWSF1c0Ez0

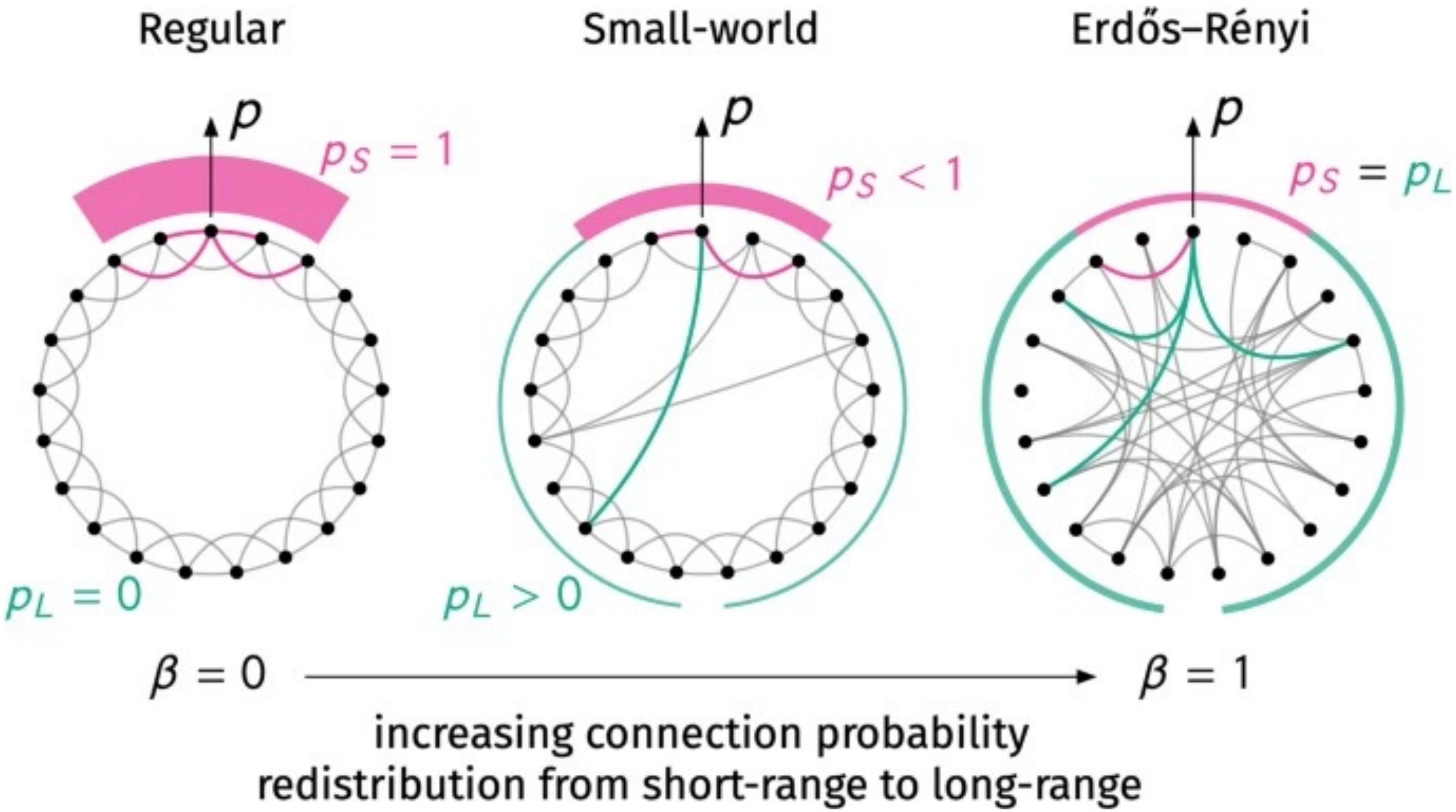

Milgram, S. (1967). The small world problem. *Psychology today*, *2*(1), 60-67.

# Small world networks

- Most nodes are not connected to most of other nodes.
- Despite the lack of dense connection, reaching from a node to another node does not need many hops.
- With increased node size, typical distance between two nodes grow logarithmically.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, *393*(6684), 440-442.

# Regular to random networks



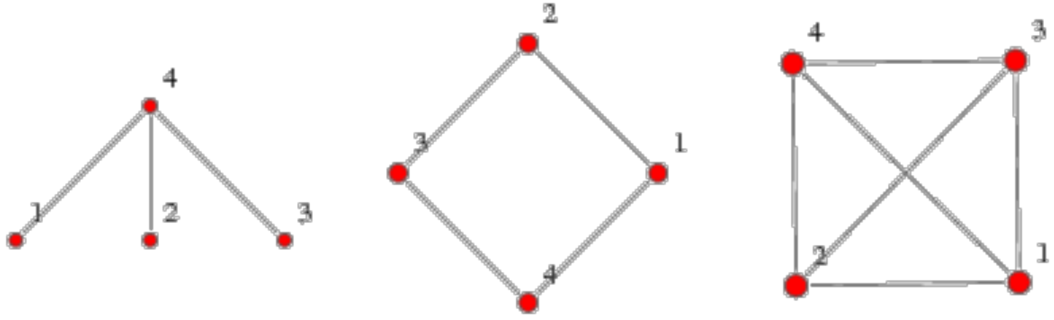

Maier, B. F. (2019). Generalization of the small-world effect on a model approaching the Erdős–Rényi random graph. *Scientific reports*, *9*(1), 1-9.

# Network degree

- Degree of a node is the count of links to that node.
- Degree of node $i$ is denoted as $k_i$ where $k_i = \sum_j a_{ij}$
  - Here $a_{ij}$ represents $i$th row and $j$th column in $A$.

# Network degree

- Degree of a node is the count of links to that node.
- Degree of node $i$ is denoted as $k_i$ where $k_i = \sum_j a_{ij}$
  - Here $a_{ij}$ represents $i$th row and $j$th column in $A$.

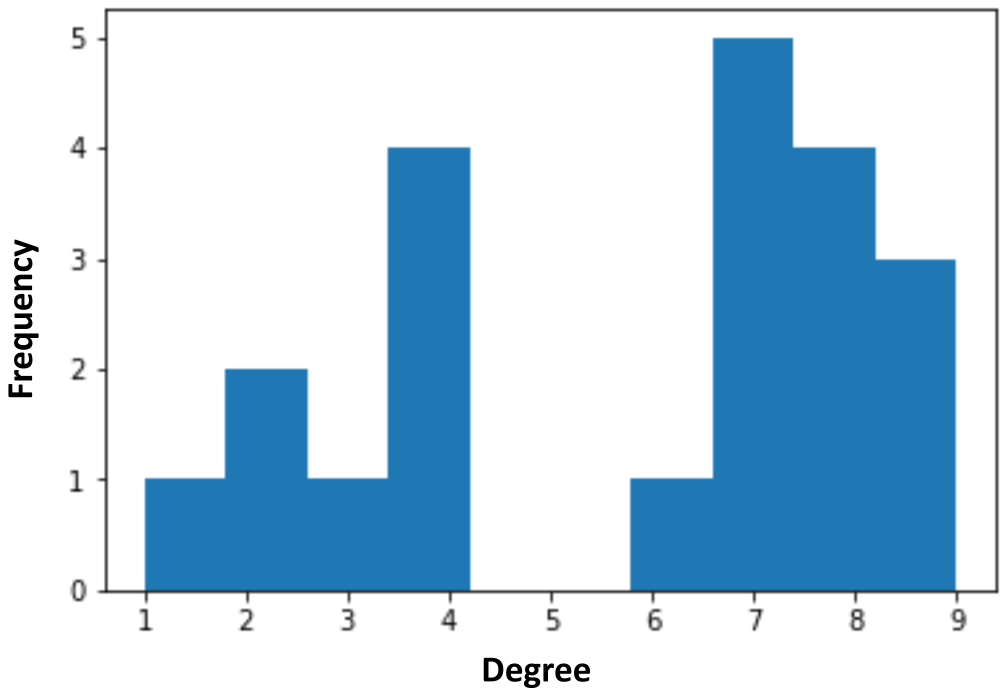



$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

Source: https://mathworld.wolfram.com/AdjacencyMatrix.html

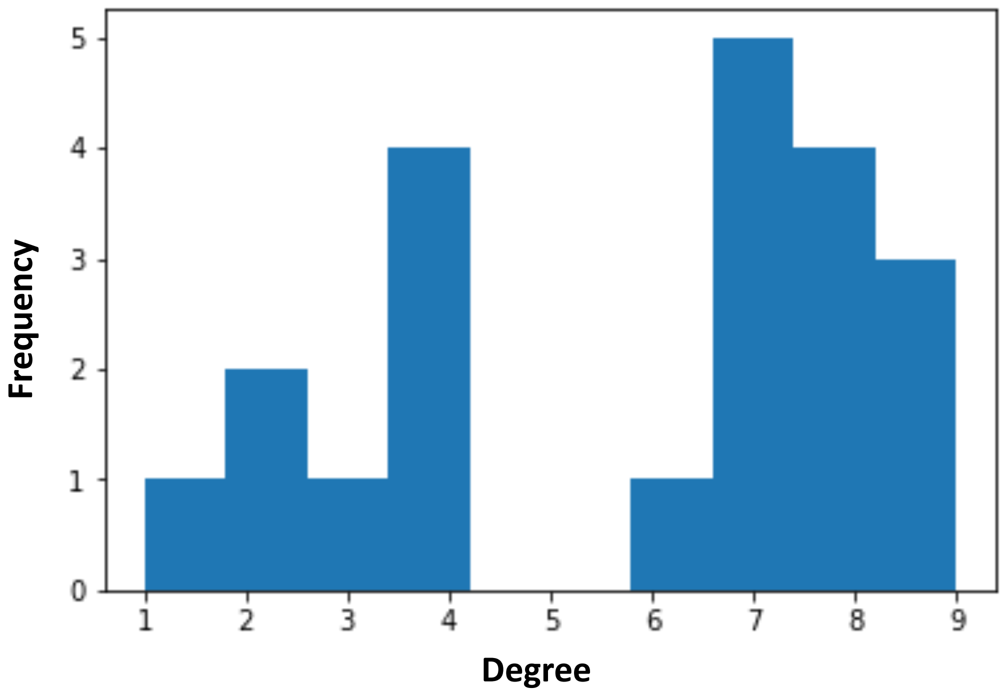
# Network degree distribution

- Captures the frequency of degrees of all nodes of the network.



# Network degree distribution

- Captures the frequency of degrees of all nodes of the network.
- One of the important structural indicators of a network.
  - Compare two networks by their degree distribution.

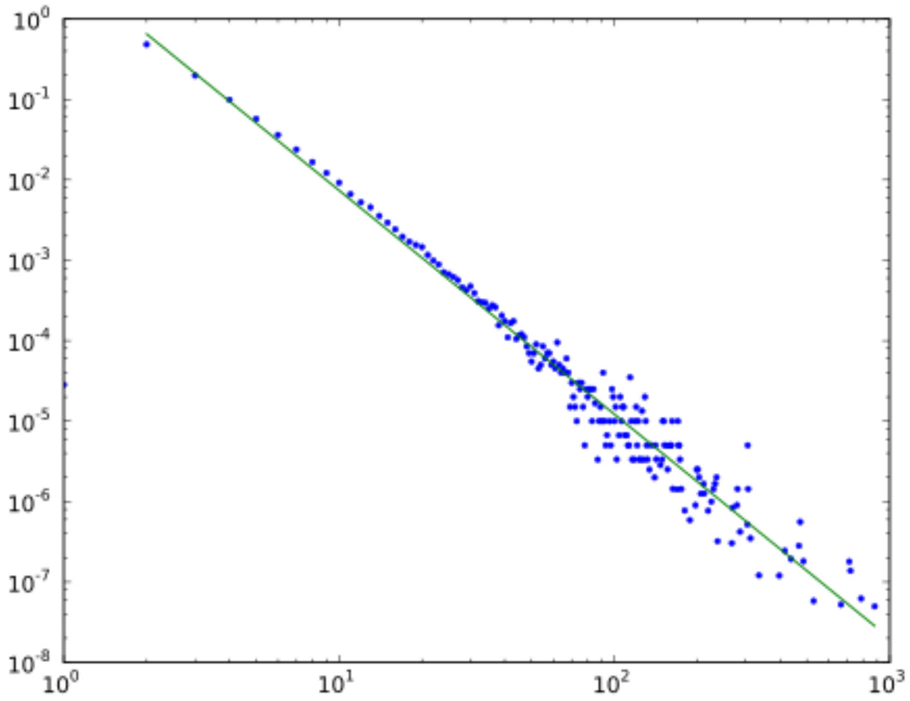


# Scale-free networks

- Degree distribution follows a power law.
- Fraction of nodes with $k$ links -> $P(k)$ where $P(k) \backsim k^{-\gamma}$ with $2 < \gamma < 3$

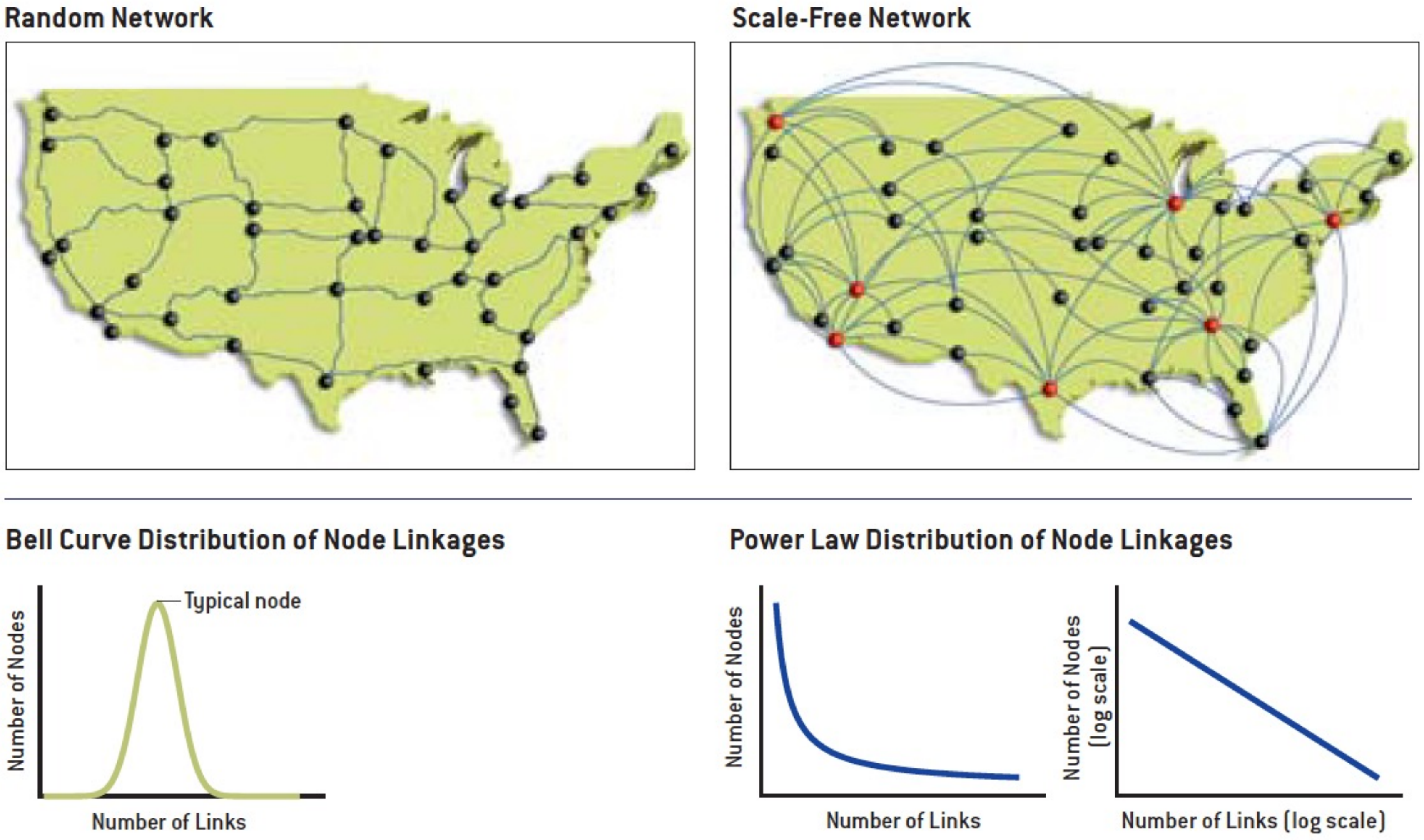
# Scale-free networks

- Degree distribution follows a power law.
- Fraction of nodes with $k$ links -> $P(k)$ where $P(k) \backsim k^{-\gamma}$ with $2 < \gamma < 3$




Source: https://commons.wikimedia.org/wiki/File:Barabasi-albert_model_degree_distribution.svg "The degree distribution of a Barabasi Albert network with 200000 nodes and with 2 new edges in each step. The exponent is -2.78 (from 2 to the max)"

# Random networks vs. scale-free networks

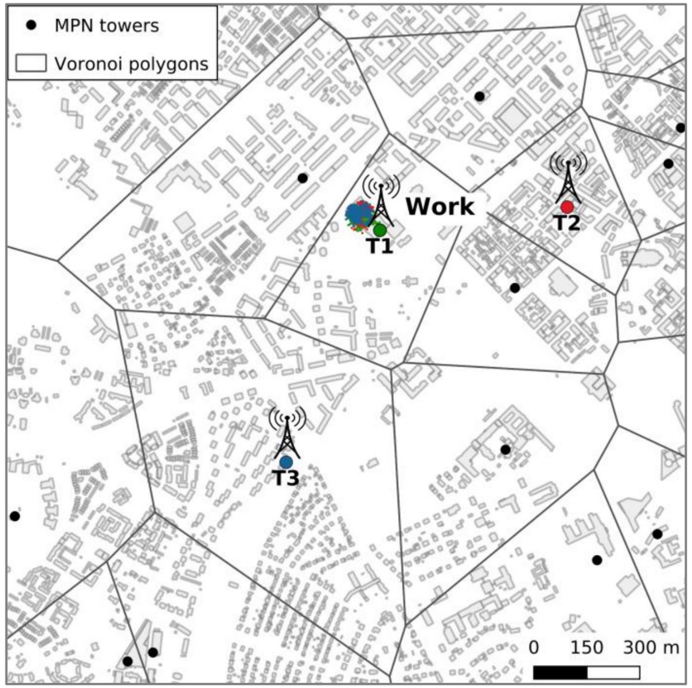



Source: Barabási, A. L., & Bonabeau, E. (2003). Scale-free networks. *Scientific american*, *288*(5), 60-69.

# Spatially-explicit networks

- A network model in which the location of nodes matter.
  - E.g.: road network, power grid network, cell network, supply chain network

GMU Campus Map




Cell tower network

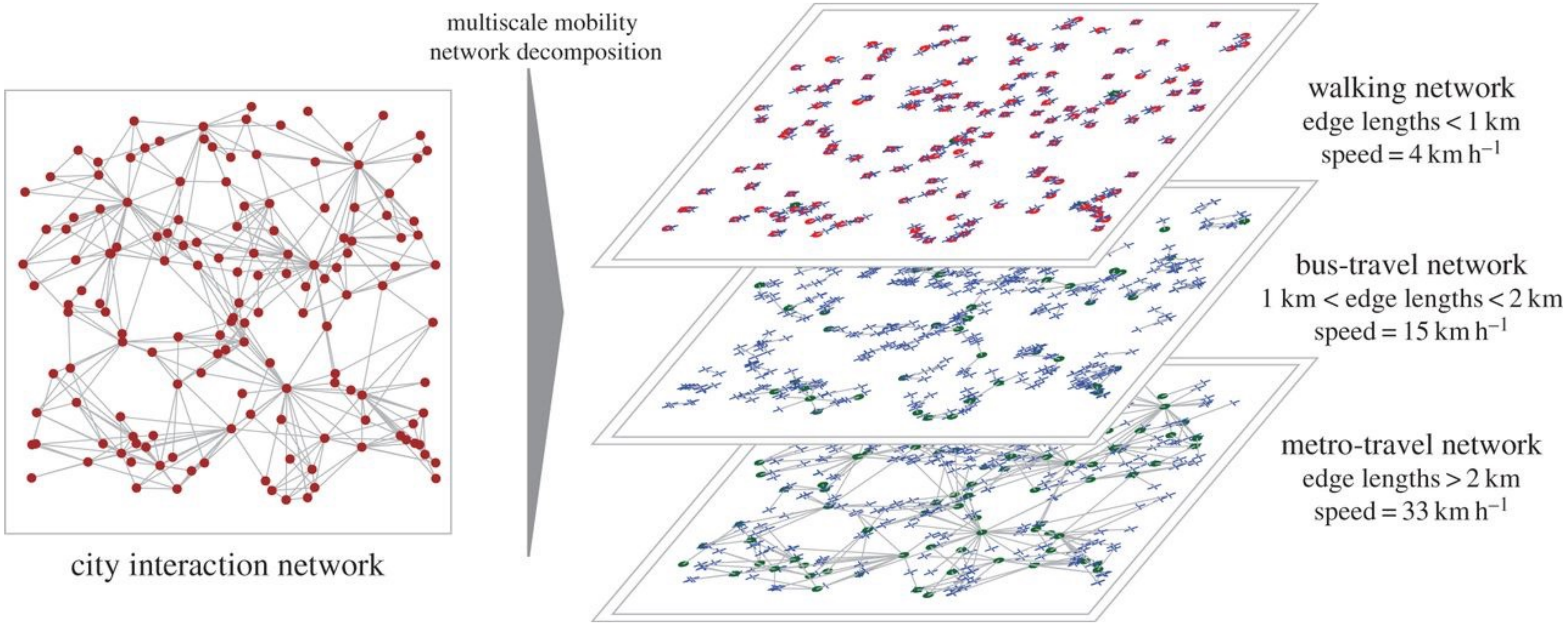
Source: Ogulenko, A., Benenson, I., Toger, M., Östh, J., & Siretskiy, A. (2021). The fallacy of the closest antenna: Towards an adequate view of device location in the mobile network. *arXiv preprint arXiv:2109.02154*.

# Nodes can have many attributes

- Location (i.e., coordinates)
- Bandwidth (if cell tower)
- Age (if human or animal)
- Gender (if human or animal)
- Type (family, workplace, school)
- …

# Multi-layer networks

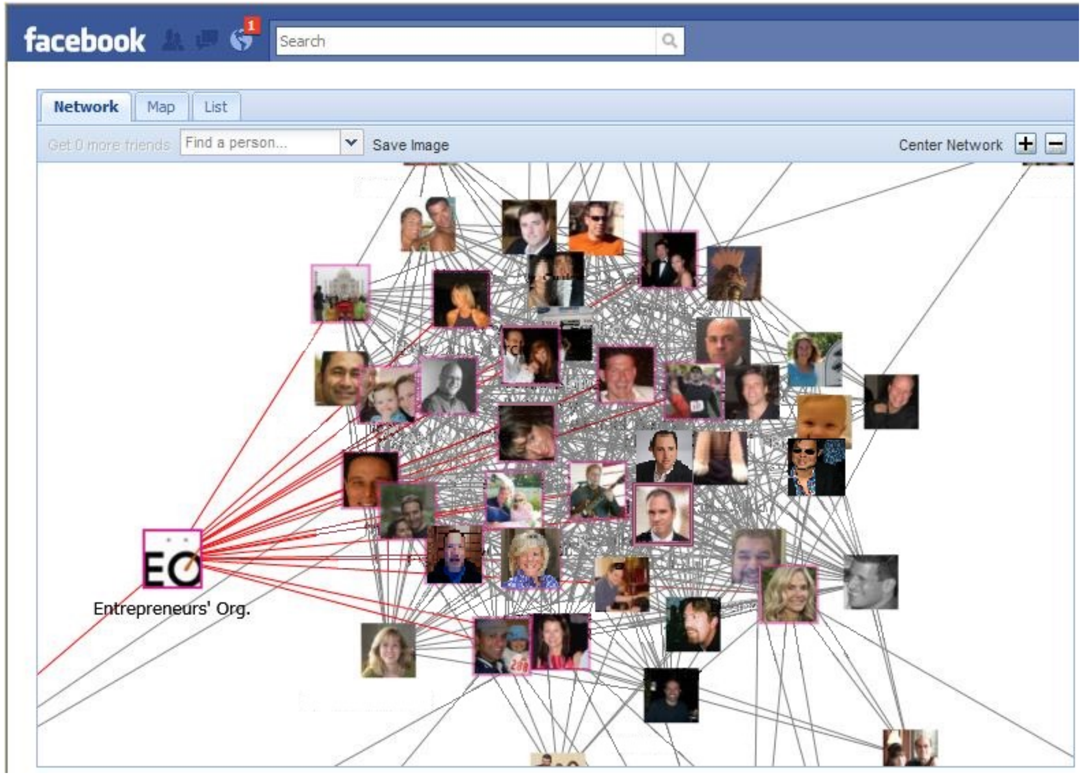
- Nodes of different types are grouped as a layer




Source: Sim, A., Yaliraki, S. N., Barahona, M., & Stumpf, M. P. (2015). Great cities look small. *Journal of The Royal Society Interface*, *12*(109), 20150315.

# Networks vs. social networks

- Social network is a specialized type of a network that captures a social phenomenon such as friendship.



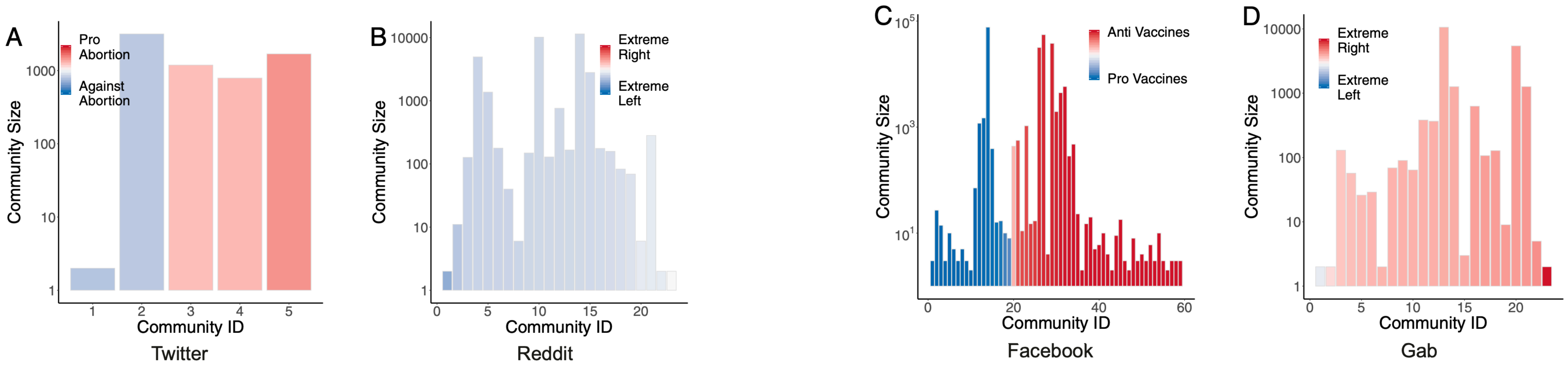

Source: https://fmsasg.com/socialnetworkanalysis/facebook/

# Online social networks

- Online social networks mostly established based on homophily, leading to echo chambers.

# Online social networks

- Online social networks mostly established based on homophily, leading to echo chambers.

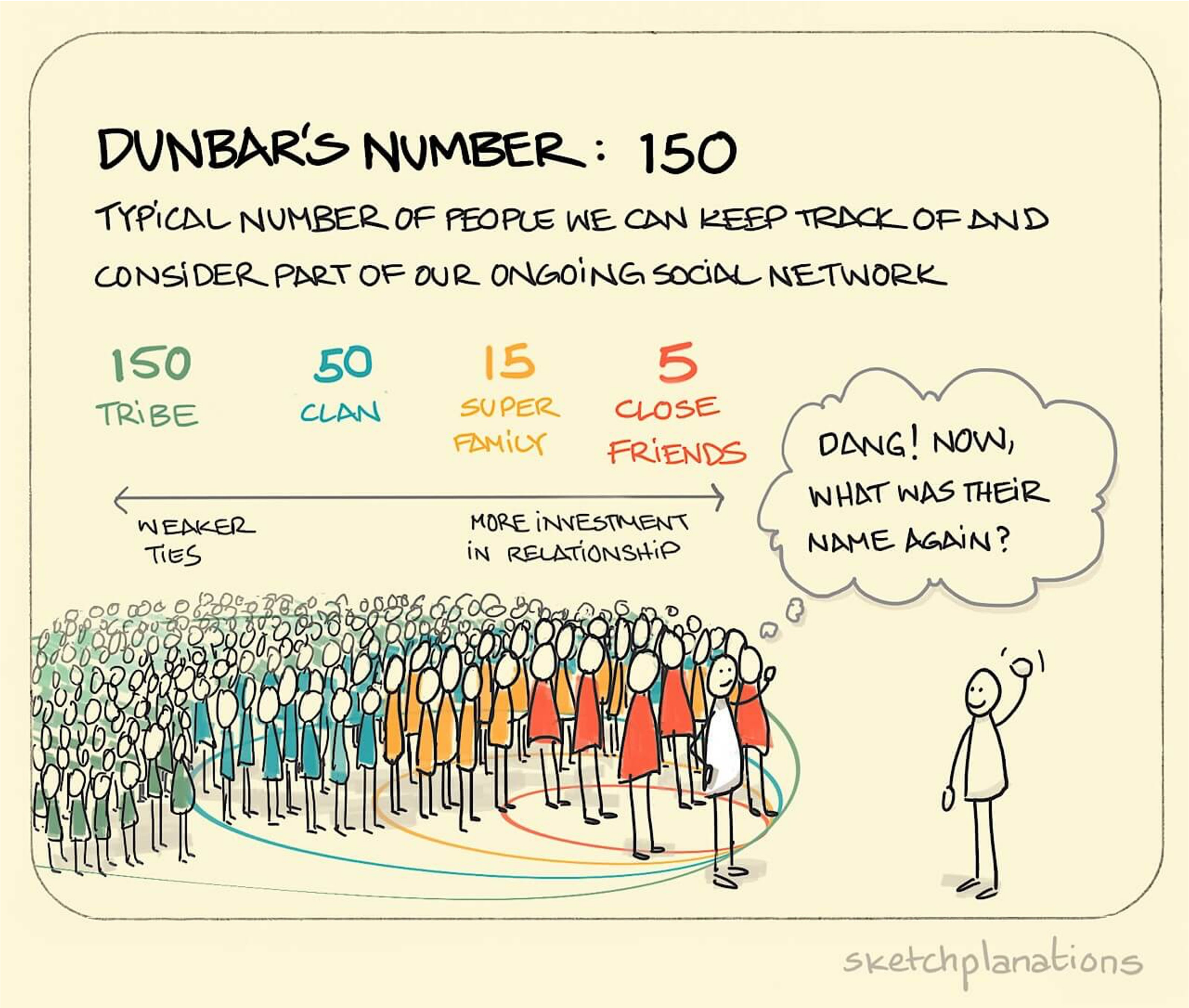



Source: Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9).

# Dunbar’s number

- Indicates a maximum number of relationship (i.e., social connections) one can handle comfortably.
- Based on average human brain size.
- Often suggested to be between 100-250 while 150 is considered typical.

Dunbar, R. I. (1992). Neocortex size as a constraint on group size in primates. *Journal of human evolution*, *22*(6), 469-493.




Source: https://sketchplanations.com/dunbars-number-150

# Community detection

- One of the common tasks in many network problems
- Grouping of similar nodes in a network (often social network)
  - Based on node features and the topology of the network
- Similar to clustering in machine learning
- No Free Lunch theorem applies

# Community detection

- Different algorithms can overfit or underfit
- Ghasemian, Hosseinmardi, & Clauset (2019) studied "16 state-of-the-art community detection algorithms applied to a novel benchmark corpus of 572 structurally diverse real-world networks"

| Algorithm | Number of Communities, $k$ | Partition Type | Link Prediction Benchmark | Link Description Benchmark | Overall Fit |
|---|---|---|---|---|---|
| Q | larger | non-probabilistic | poor | good | over fits |
| Q-MR | larger | non-probabilistic | poor | good | over fits |
| Q-MP | smaller | spectral/ non-probabilistic | poor | poor | under fits |
| Q-GMP | smaller | spectral/ non-probabilistic | — | — | inconclusive |
| B-NR (SBM) | smaller | probabilistic | very good | moderate | well-fitted |
| B-NR (DC-SBM) | smaller | probabilistic | moderate | very good | over fits, modestly |
| B-HKK (SBM) | smaller | probabilistic | good | moderate | under fits, modestly |
| cICL-HKK (SBM) | smaller | probabilistic | good | moderate | under fits, modestly |
| Infomap | larger | non-probabilistic | moderate | moderate | over fits |
| MDL (SBM) | smaller | probabilistic | good | poor | under fits |
| MDL (DC-SBM) | smaller | probabilistic | very good | moderate | well-fitted |
| S-NB | smaller | spectral | moderate | moderate | uneven fits |
| S-cBHm | smaller | spectral | — | — | inconclusive |
| S-cBHa | smaller | spectral | — | — | inconclusive |
| AMOS | larger | non-probabilistic | — | — | inconclusive |
| LRT-WB (DC-SBM) | larger | non-probabilistic | — | — | inconclusive |

Source: Ghasemian, A., Hosseinmardi, H., & Clauset, A. (2019). Evaluating overfit and underfit in models of network community structure. *IEEE Transactions on Knowledge and Data Engineering*, *32*(9), 1722-1735.

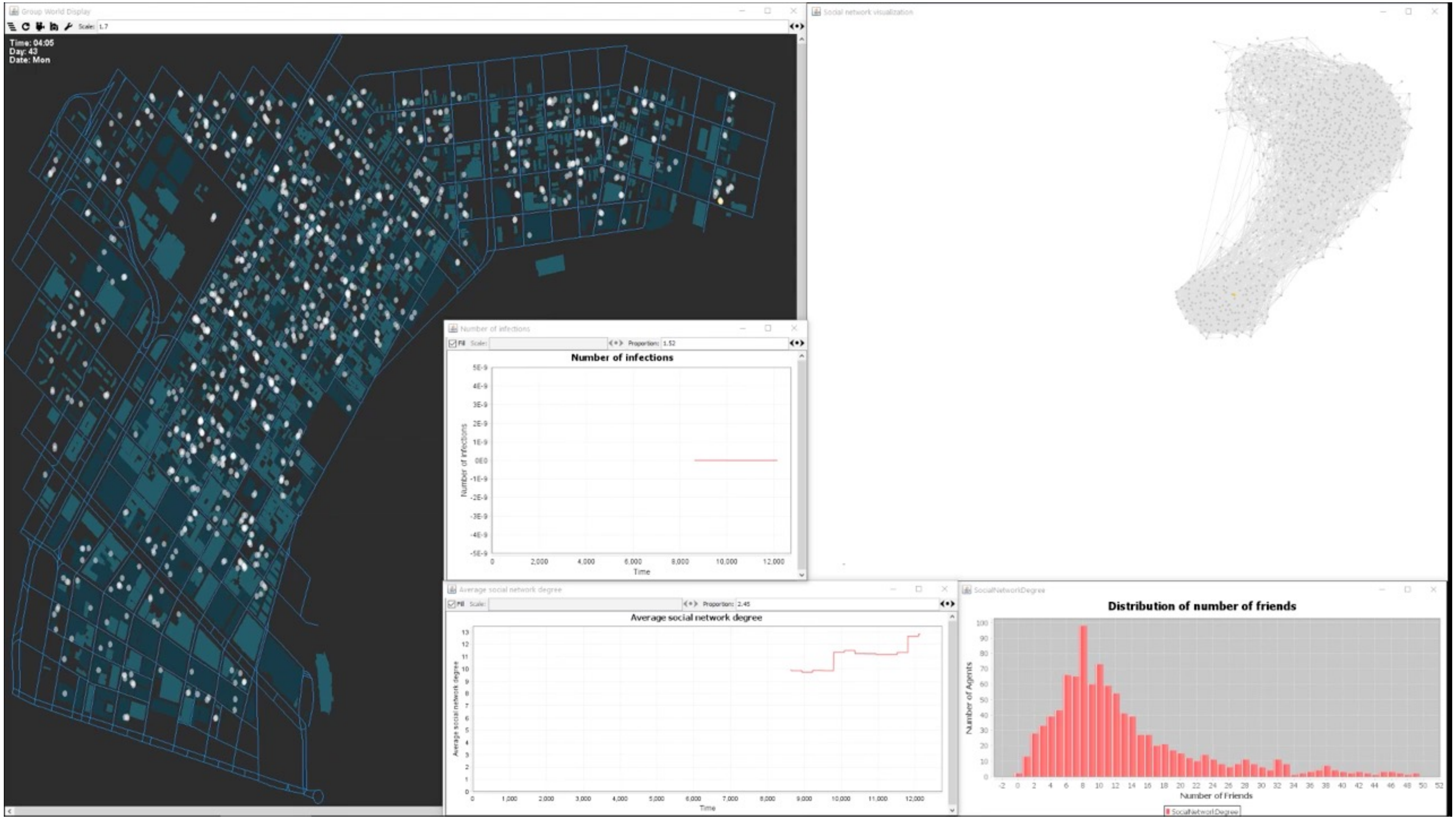
# Identifying central nodes

- Several measures are proposed
  - Degree centrality
  - Closeness centrality
  - Betweenness centrality
  - PageRank
  - …
- Has applications to identify the most influential node.
  - Who is the best person to propagate our message?

# Network generation

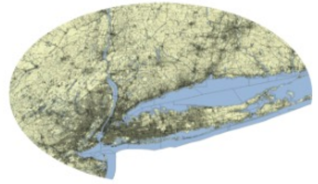
- Rule-based network generation
  - E.g.: Preferential attachment
  - Advantage: simple to implement
  - Disadvantage: time notion and detail
- Agent-based network generation
  - E.g.: Co-location-based
  - Advantage: more grounded in real world and can capture time
  - Disadvantage: complex to implement

# Example model: French Quarter epidemic




Source: https://www.youtube.com/watch?v=3pBdnT8a9LA

# Synthetic population generation

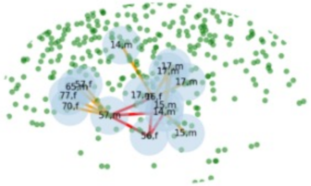


**Step 1**

| Create Individuals and Space | | |
|---|---|---|
| **Tasks** | **Data** | **Python Package** |
| • Create Living Space for Home, Work Locations on Road Networks | Road Network | Geopandas, Pandas, Numpy |
| • Create Individuals | Census Data 2010 | Pandas, Random |
| • Group Individuals into Household | Census Data 2010 | Pandas, Itertools, Numpy, Random |

**Step 2**

| Assign Daytime Locations | | |
|---|---|---|
| **Tasks** | **Data** | **Python Package** |
| • Create Workplaces and Assign them to Individuals | Data Generated from Step 1, LEHDs LODES | Pandas, Sklearn, Numpy, Random |
| • Place Home and Work Places on the Spaces | Road Network | Pandas, Geopandas |
| • Assign Educational Sites based on age and the closest distance | Data Generated from Step 1, EPA OEI | Pandas, Geopandas |



**Step 3**

| Create Social Networks | | |
|---|---|---|
| **Tasks** | **Data** | **Python Package** |
| • Create 3 types of Social Networks (e.g., Household, Work, and Educational) | Data Generated from Step 1 & 2 | Pandas, Networkx |

Source: Jiang, N., Kavak, H., Kennedy, W. G., & Crooks, A. T. (2021). Computational Urban Science (Under review)