CS5056
DATA ANALYTICS
GRAPH THEORY

HÉCTOR G. CEBALLOS, FRANCISCO J. CANTÚ

CEBALLOS@TEC.MX, FCANTU@TEC.MX

Image: https://towardsdatascience.com/graph-theory-on-to-network-theory-379b390fb19b

GRAPH THEORY

- Graph theory is the study of graphs, which are mathematical structures used to model *pairwise relations* between objects.
- A graph is made up of vertices (also called nodes or points) which are connected by edges (also called links or lines).
  - Undirected graphs: edges link two vertices symmetrically
  - Directed graphs: edges link two vertices asymmetrically

DIRECTED VS INDIRECTED GRAPH

- If the vertices represent people at a party, and there is an edge between two people if they shake hands, then this graph is _____ because any person *A* can shake hands with a person *B* only if *B* also shakes hands with *A*.
- If any edge from a person *A* to a person *B* corresponds to *A* owes money to *B*, then this graph is _____, because owing money is not necessarily reciprocated.

GRAPH

- A <u>graph</u> is a pair $G = (V, E)$, where V is a set of *vertices* and E is a set of two-sets (sets with two distinct elements) of vertices ($V \times V$), i.e. edges.
  - The vertices x and y of an edge $\{x, y\}$ are called the *endpoints* of the edge.
  - The edge is said to *join* x and y and to be *incident* on x and y.
  - A vertex may not belong to any edge.
- A <u>multigraph</u> is a generalization that allows multiple edges to have the same pair of endpoints.

GRAPH

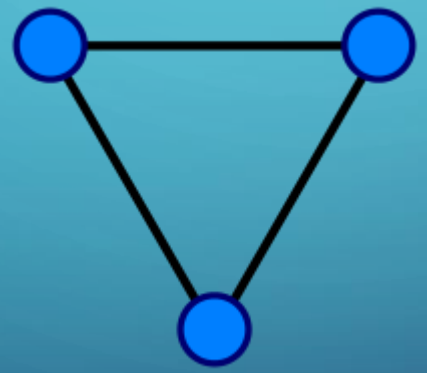
- An empty graph is a graph that has an empty set of vertices (and thus an empty set of edges).
- Sometimes, graphs are allowed to contain loops, which are edges that join a vertex to itself.
- The order of a graph is its number of vertices $|V|$.
- The size of a graph is its number of edges $|E|$.
- In a graph of order n, the maximum degree of each vertex is $n - 1$ (or n if loops are allowed), and the maximum number of edges is $n(n - 1)/2$ (or $n(n + 1)/2$ if loops are allowed).

GRAPH DATA STRUCTURES

- The edges of a graph define a symmetric relation on the vertices, called the *adjacency relation*. Specifically, two vertices x and y are *adjacent* if $\{x, y\}$ is an edge.
- A graph may be fully specified by its adjacency matrix A, which is an nxn square matrix, with A_{ij} specifying the nature of the connection between vertex i and vertex j.
- For a simple graph, $A_{ij}= 0$ or 1, indicating disconnection or connection respectively, with $A_{ii}=0$.
- Undirected graphs will have a symmetric adjacency matrix ($A_{ij}=A_{ji}$).

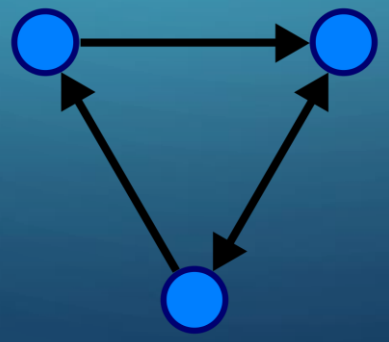
GRAPH VISUALIZATION

- Typically, a graph is depicted in diagrammatic form as a set of dots or circles for the vertices, joined by lines or curves for the edges.



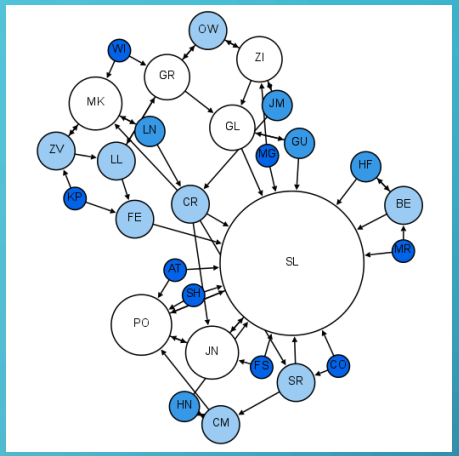
DIRECTED GRAPH

- A **directed graph** or **digraph** is a graph in which edges have orientations.
- It is an ordered pair {\displaystyle G=(V,E)}G=(V,E) comprising:
  - V, a set of vertices (also called nodes or points);
  - $E \subseteq \{(x,y)|(x,y) \in V^2 \; and \; x \neq y\}$, a set of edges (also called directed edges, directed links, directed lines, arrows or arcs) which are ordered pairs of vertices (that is, an edge is associated with two distinct vertices).



APPLICATIONS
SOCIAL SCIENCES

Sociogram



- Widely used in sociology as a way, for example, to measure actors' prestige or to explore rumor spreading, notably through the use of Social Network Analysis (SNA) software.
  - **Acquaintanceship and friendship graphs** describe whether people know each other.
  - **Influence graphs** model whether certain people can influence the behavior of others.
  - **Collaboration graphs** model whether two people work together in a particular way, such as acting in a movie together.

SOCIAL NETWORK ANALYSIS (SNA)

NETWORK FEATURES

- Density: ratio of the number of edges E to the number of possible edges in a network with N nodes.
- Average Degree: the average degree for all nodes.
- Diameter: the longest of all the calculated shortest paths in a network.
- Clustering coefficient: measure of an "all-my-friends-know-each-other".

The clustering coefficient of the i'th node is

$$C_i = \frac{2e_i}{k_i(k_i - 1)},$$

where k_i is the number of neighbours of the i'th node, and e_i is the number of connections between these neighbours. The maximum possible number of connections between neighbors is, then,

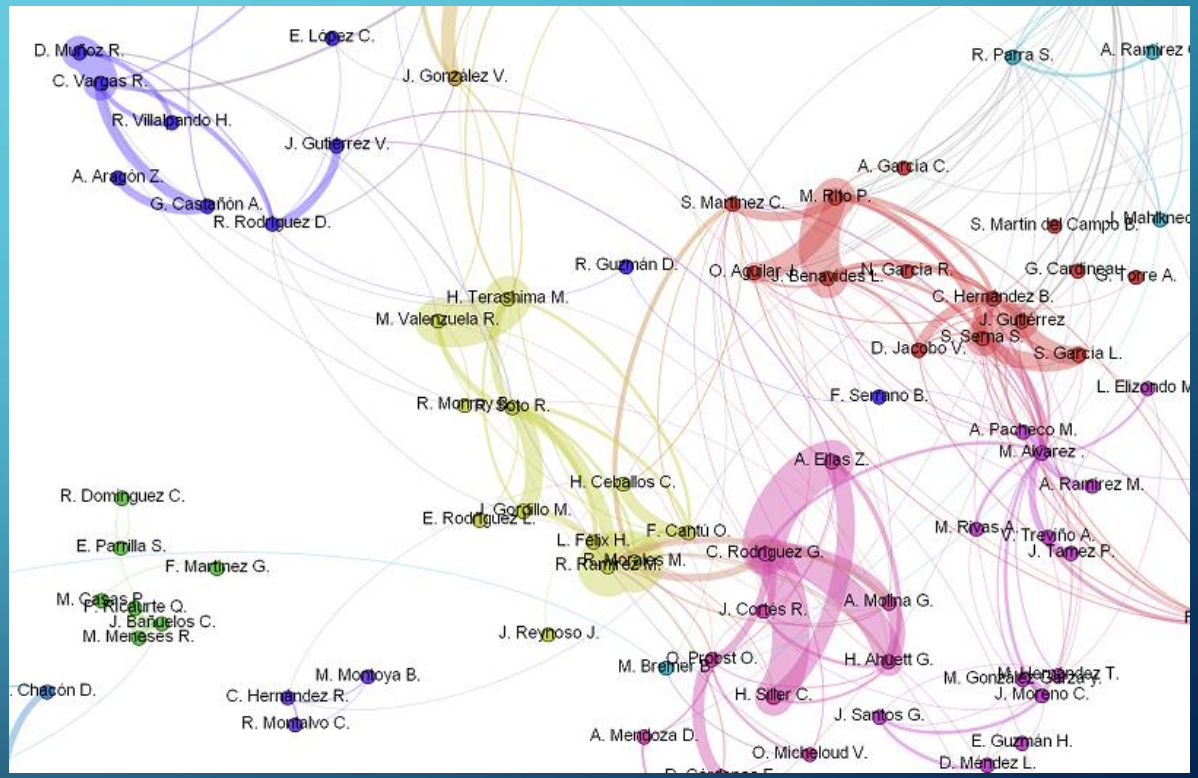
$$\binom{k}{2} = \frac{k(k-1)}{2}.$$

COAUTHORSHIP NETWORKS

- Nodes: Authors
- Edges: # coauthored papers (similarity)

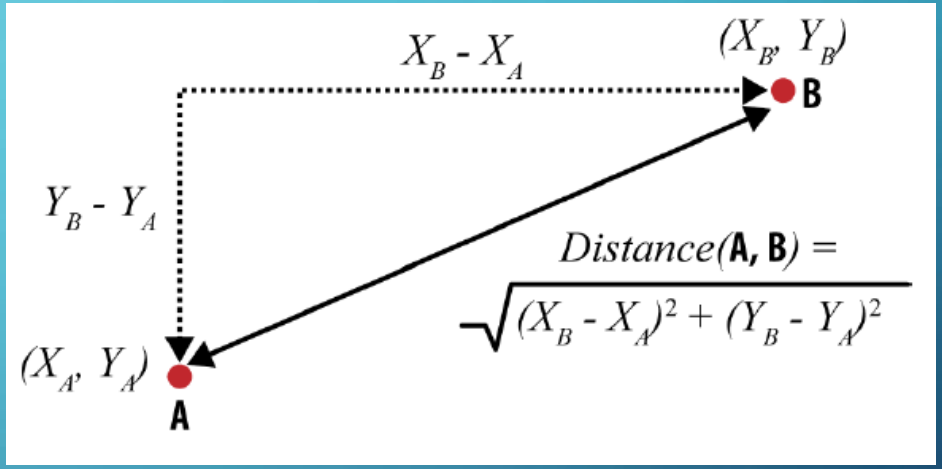
Answer

1. Who are author neighbors?
2. What does it represent an author cluster?
3. Who are similar authors?
4. Which recommendations could you make?



SIMILARITY AND DISTANCE

- Objects are represented by vectors of features.
- Similarity is estimated as the distance between every pair of objects.



Answer

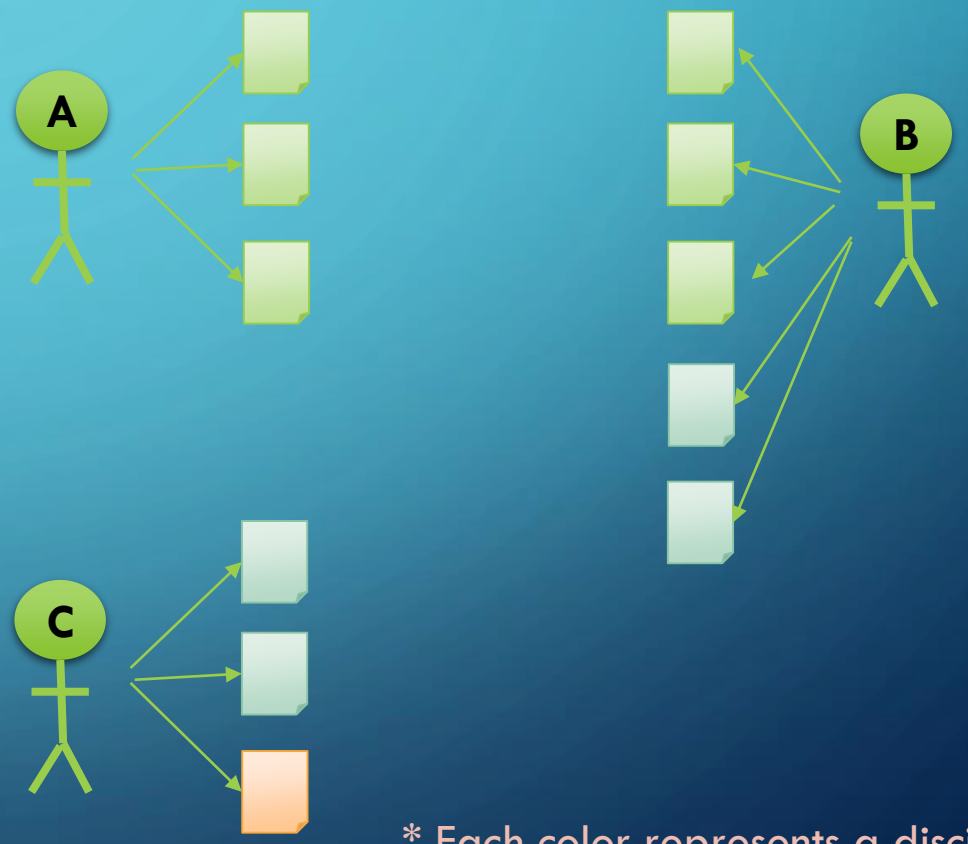
- Which features would you consider for identifying *similar* authors?

Equation 6-1. General Euclidean distance

$$\sqrt{(d_{1,A} - d_{1,B})^2 + (d_{2,A} - d_{2,B})^2 + \ldots + (d_{n,A} - d_{n,B})^2}$$

NEAREST-NEIGHBOR REASONING

- **Problem:** Identify authors working on a similar research topic (match-making)?



Answer

- Which features would you choose to represent each author?
- Which distance would you use?

* Each color represents a discipline.

ISSUES WITH NEAREST-NEIGHBOR METHODS

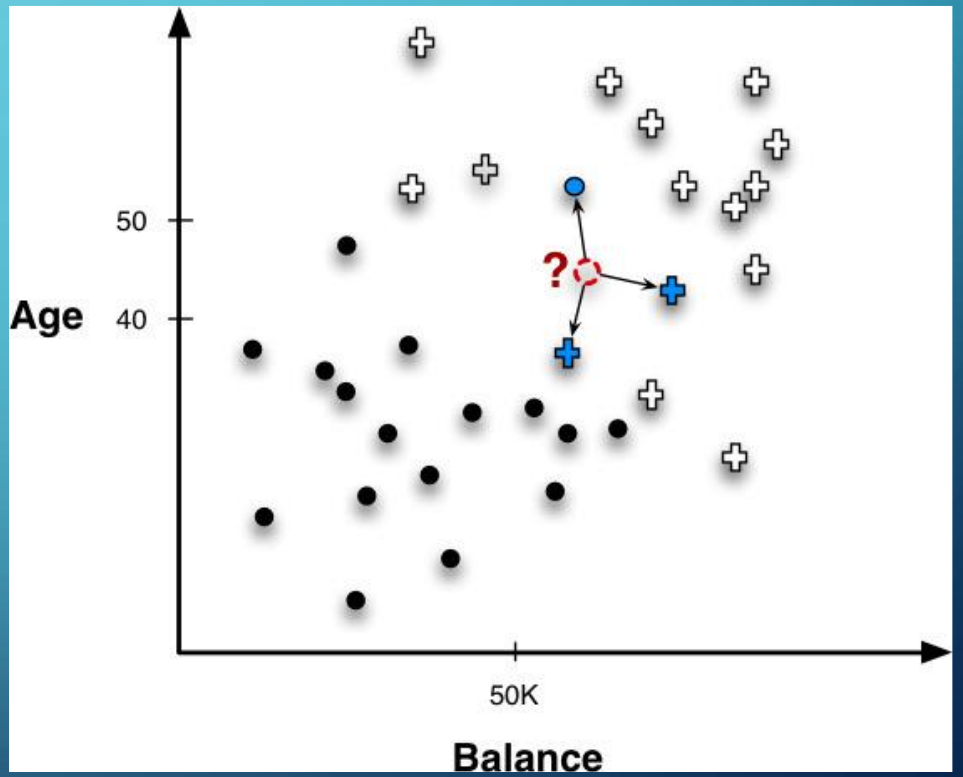
- **Intelligibility:** the justification of a specific *decision* and the intelligibility of an entire *model*.
- **Dimensionality and domain knowledge:** multiple ranges of numeric attributes; the effect of one attribute must be properly scaled to avoid swamp the effect of another with a much smaller range.
- **Computational Efficiency:** Cross product in the worst scenario.

Answer

- How would you justify collaboration to a researcher based on topic similarity?
- How could you estimate *expertise* of a researcher in a topic?
- How would you prune the topic similarity network?

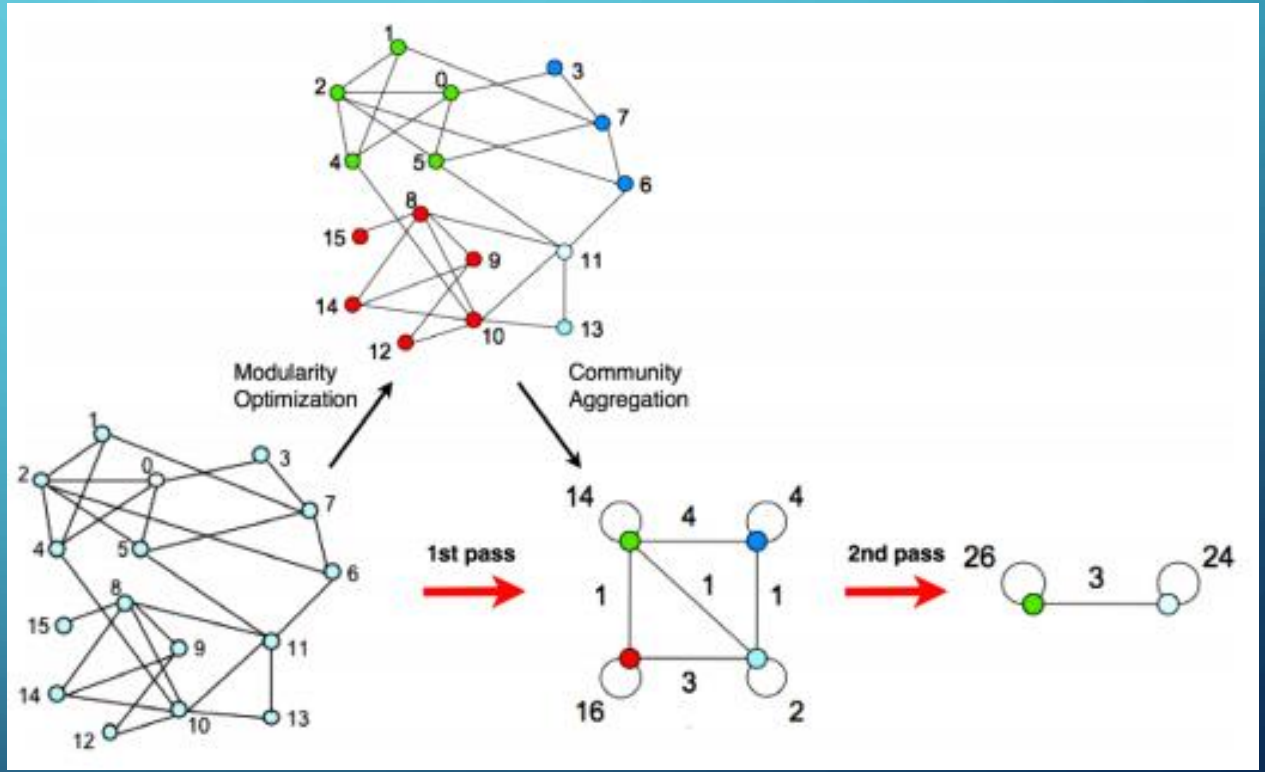
NEAREST NEIGHBORS FOR PREDICTIVE MODELING

- ***Classification***: *voting of nearest neightbors.*
  - *Is author A working in BIO?*
- ***Probability estimation***: *score rather than a class.*
  - *How likely is that author A collaborates with author B?*
- ***Regression***: *estimate missing value.*
  - *How likely is that Tec collaborate with Fudan University in BIO?*



HIERARCHICAL CLUSTERING (SOCIAL NETWORKS)

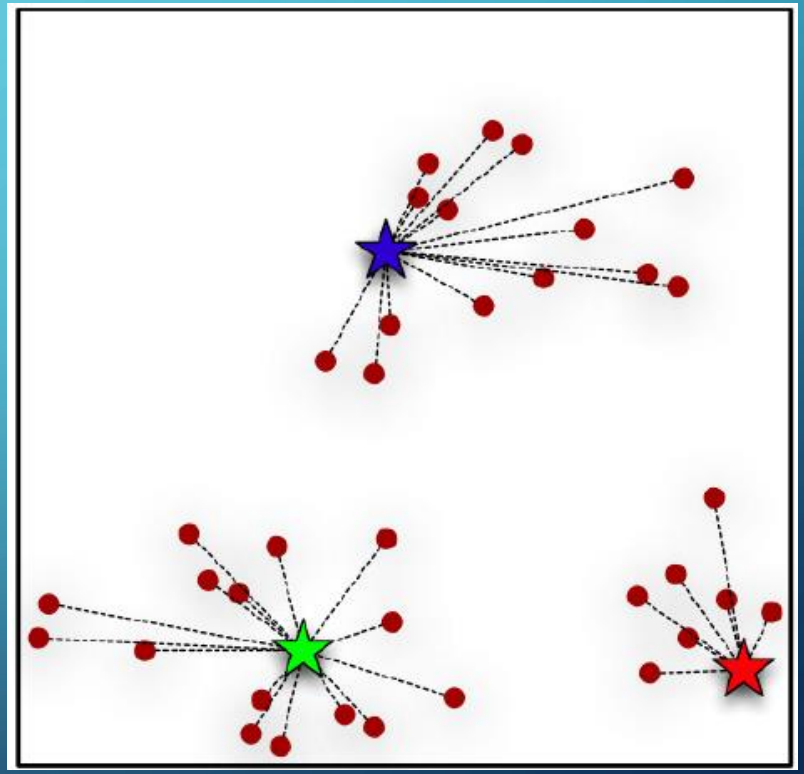
Blondel, Vincent D; Guillaume, Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne (9 October 2008). "**Fast unfolding of communities in large networks**". *Journal of Statistical Mechanics: Theory and Experiment*. **2008** (10): P10008. arXiv:0803.0476. Bibcode:2008JSMTE..10..008B. doi:10.1088/1742-5468/2008/10/P10008.



CLUSTERING AROUND CENTROIDS

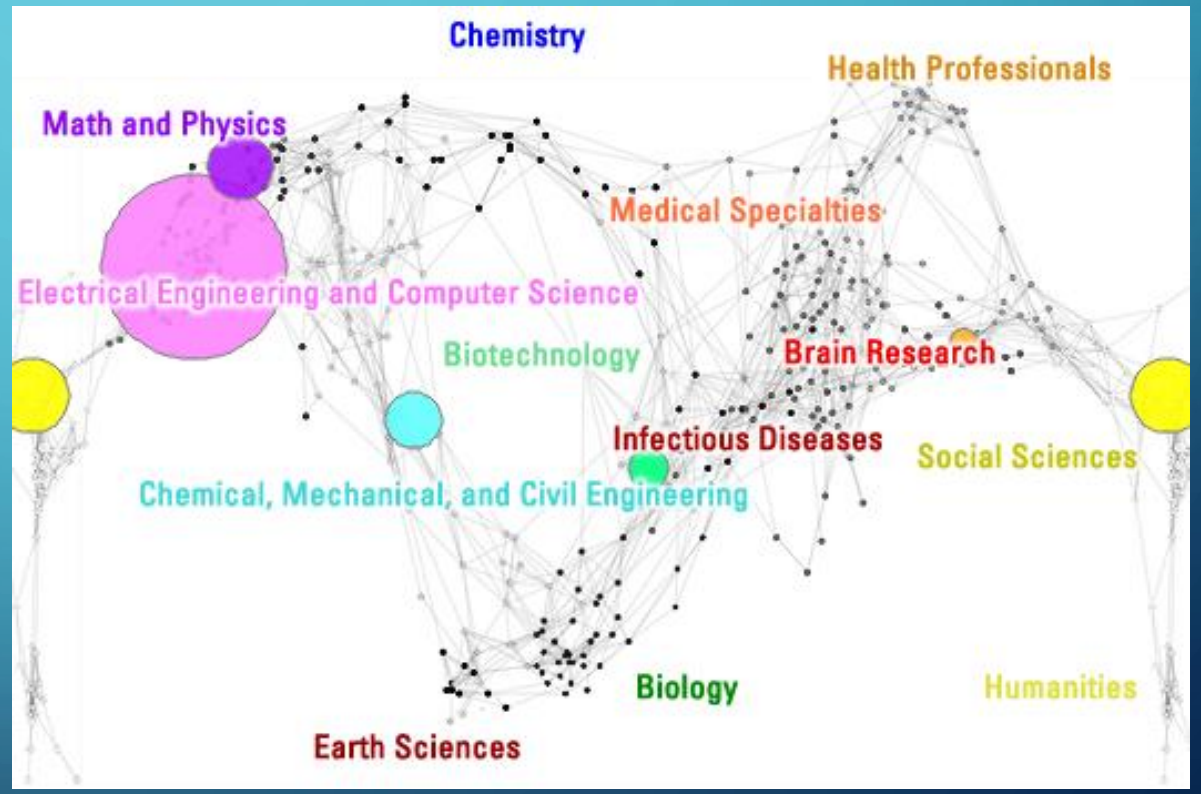
Informed vs random

- How would you define centroids in a *coauthorship* network?
- How would you define centroids in a *topic similarity* network?



UNDERSTANDING CLUSTERING

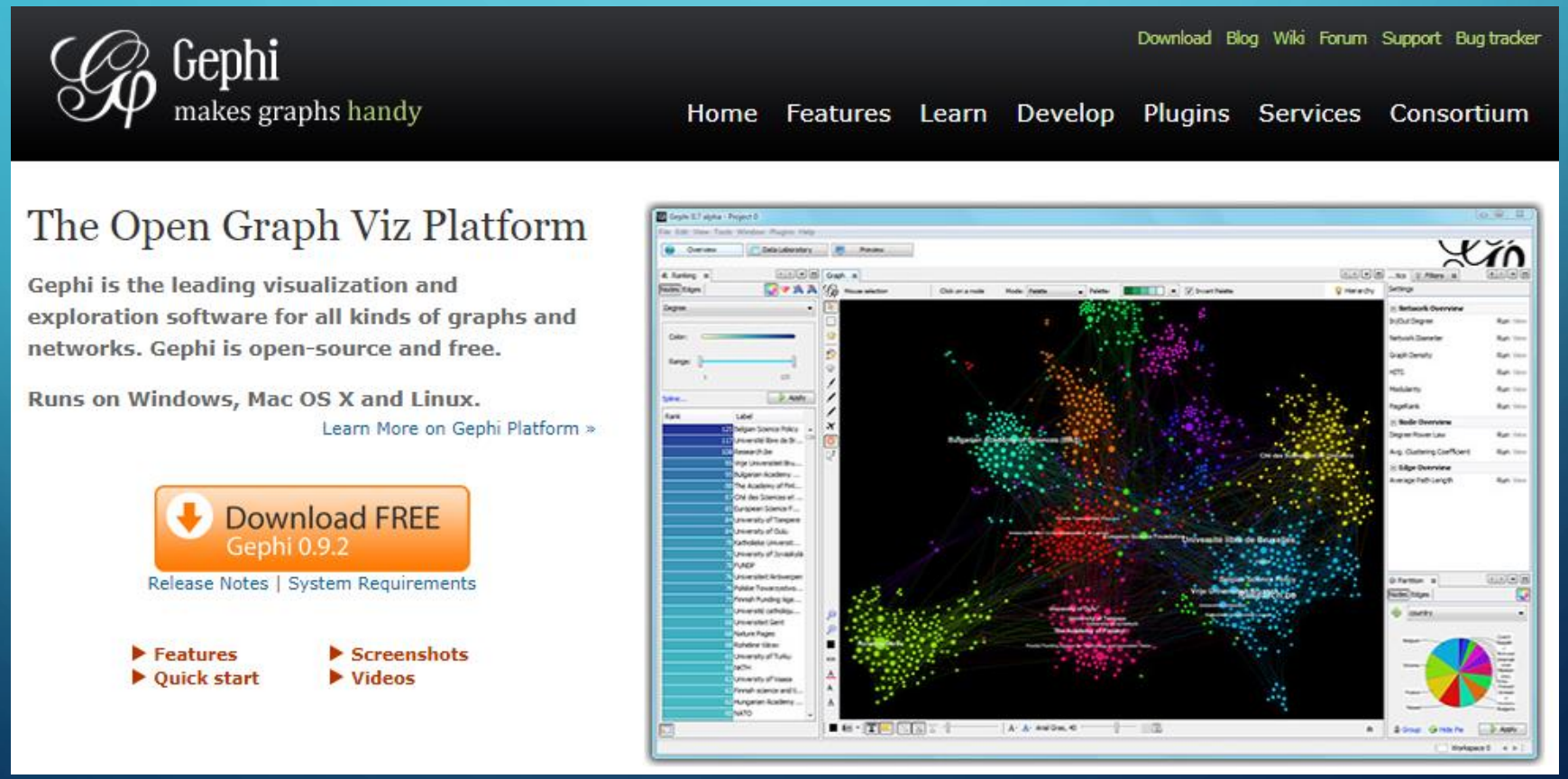
- Map of Science of the Intelligent Systems group
  - Nodes represent disciplines.
  - Related disciplines are linked in a force lines layout.
  - Diameter of nodes represent the number of papers published.
- What information this map provides to a person looking for research collaboration?



GEPHI TUTORIAL

DOWNLOAD GEPHI

- https://gephi.org/



GEFX

- XML format
  - Attributes
  - Nodes
  - Edges (weight)

```xml
<?xml version="1.0" encoding="UTF-8"?>
<gexf xmlns="http://www.gexf.net/1.2draft" xmlns:xsi="http://www.w3.
    <meta lastmodifieddate="2009-03-20">
        <creator>Gephi.org</creator>
        <description>A Web network</description>
    </meta>
    <graph defaultedgetype="directed">
        <attributes class="node">
            <attribute id="0" title="url" type="string"/>
            <attribute id="1" title="indegree" type="float"/>
            <attribute id="2" title="frog" type="boolean">
                <default>true</default>
            </attribute>
        </attributes>
        <nodes>
            <node id="0" label="Gephi">
                <attvalues>
                    <attvalue for="0" value="http://gephi.org"/>
                    <attvalue for="1" value="1"/>
                </attvalues>
            </node>
                                .
                                .
                                .
            <node id="3" label="BarabasiLab">
                <attvalues>
                    <attvalue for="0" value="http://barabasilab.com"
                    <attvalue for="1" value="1"/>
                    <attvalue for="2" value="false"/>
                </attvalues>
            </node>
        </nodes>
        <edges>
            <edge id="0" source="0" target="1"/>
            <edge id="1" source="0" target="2"/>
            <edge id="2" source="1" target="0"/>
            <edge id="3" source="2" target="1"/>
            <edge id="4" source="0" target="3"/>
        </edges>
    </graph>
</gexf>
```

DEMO

- Network: Topic Similarity EIC-Chinesse Researchers
  - 530 Nodes , 19683 Edges
  - Researcher features: Research Group, Role, Topics (AI, Bioengineering, etc.), School/University.