Applications

Of Data Science

Introduction to Networks

Applications of Data Science - Class 6

Giora Simchoni

gsimchoni@gmail.com and add #dsapps in subject

Stat. and OR Department, TAU

2020-01-18

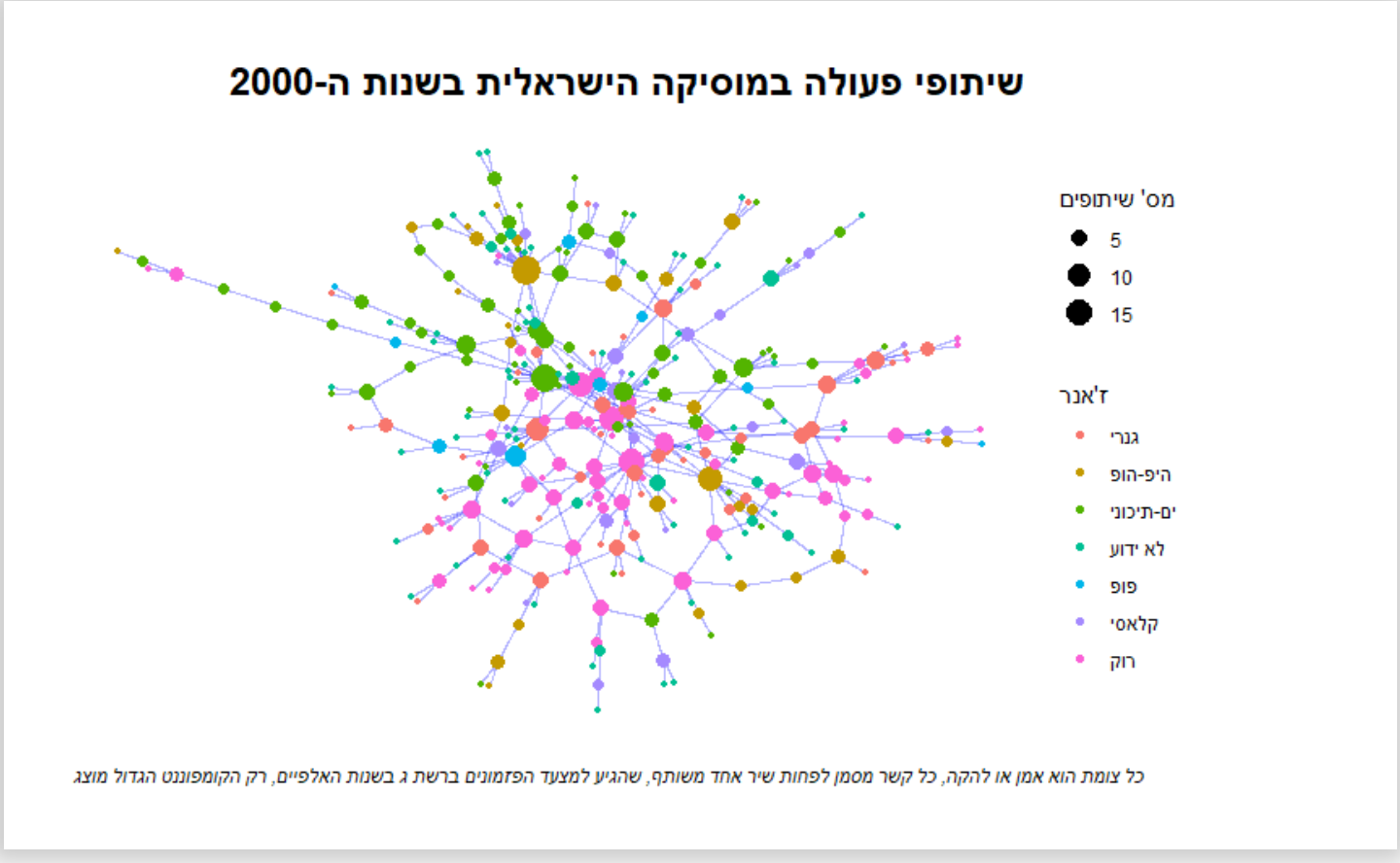
Why Networks?

Because this

	A	B	C	D
1	url	loc	artist	song
2	http://pizmonet.co.il/wiki/2.1.2000	1	אסף אמדורסקי	חלום נהה
3	http://pizmonet.co.il/wiki/2.1.2000	2	דנה אינטרנשיונל	עד סוף הזמן
4	http://pizmonet.co.il/wiki/2.1.2000	3	רוקפור ואביב גפן	שוב לא שקט
5	http://pizmonet.co.il/wiki/2.1.2000	4	יזהר אשדות וברי סחרוף	שה לה לה
6	http://pizmonet.co.il/wiki/2.1.2000	5	מיכל אמדורסקי	אין לי מה לומר
7	http://pizmonet.co.il/wiki/2.1.2000	6	יזהר אשדות וברי סחרוף	זמן קסם
8	http://pizmonet.co.il/wiki/2.1.2000	7	אסף אמדורסקי	15 דקות
9	http://pizmonet.co.il/wiki/2.1.2000	8	אביתר בנאי	חלון
10	http://pizmonet.co.il/wiki/2.1.2000	9	מיקה קרני	אלול בין כרם
11	http://pizmonet.co.il/wiki/2.1.2000	10	רוקפור	הבעס
12	http://pizmonet.co.il/wiki/2.1.2000	11	מיקה קרני	תגיד עבשיו
13	http://pizmonet.co.il/wiki/2.1.2000	12	החשמליות	לונדון
14	http://pizmonet.co.il/wiki/2.1.2000	13	שב"ק ס'	הנה זה בא
15	http://pizmonet.co.il/wiki/2.1.2000	14	מיכל אמדורסקי	אל תלך עבשיו
16	http://pizmonet.co.il/wiki/2.1.2000	15	אלי לוזון	הכול בשבילך
17	http://pizmonet.co.il/wiki/2.1.2000	16	עברי לידר	הכוס הכחולה
18	http://pizmonet.co.il/wiki/2.1.2000	17	בתי	קצת ממך
19	http://pizmonet.co.il/wiki/2.1.2000	18	ריטה	כמה אירוני

Can only get you so far.

Divided they sing



Networks Overview

A Network

A network, is comprised of:

- Nodes (vertices, points, actors), joined together in pairs by
- Edges (links, connections, ties)

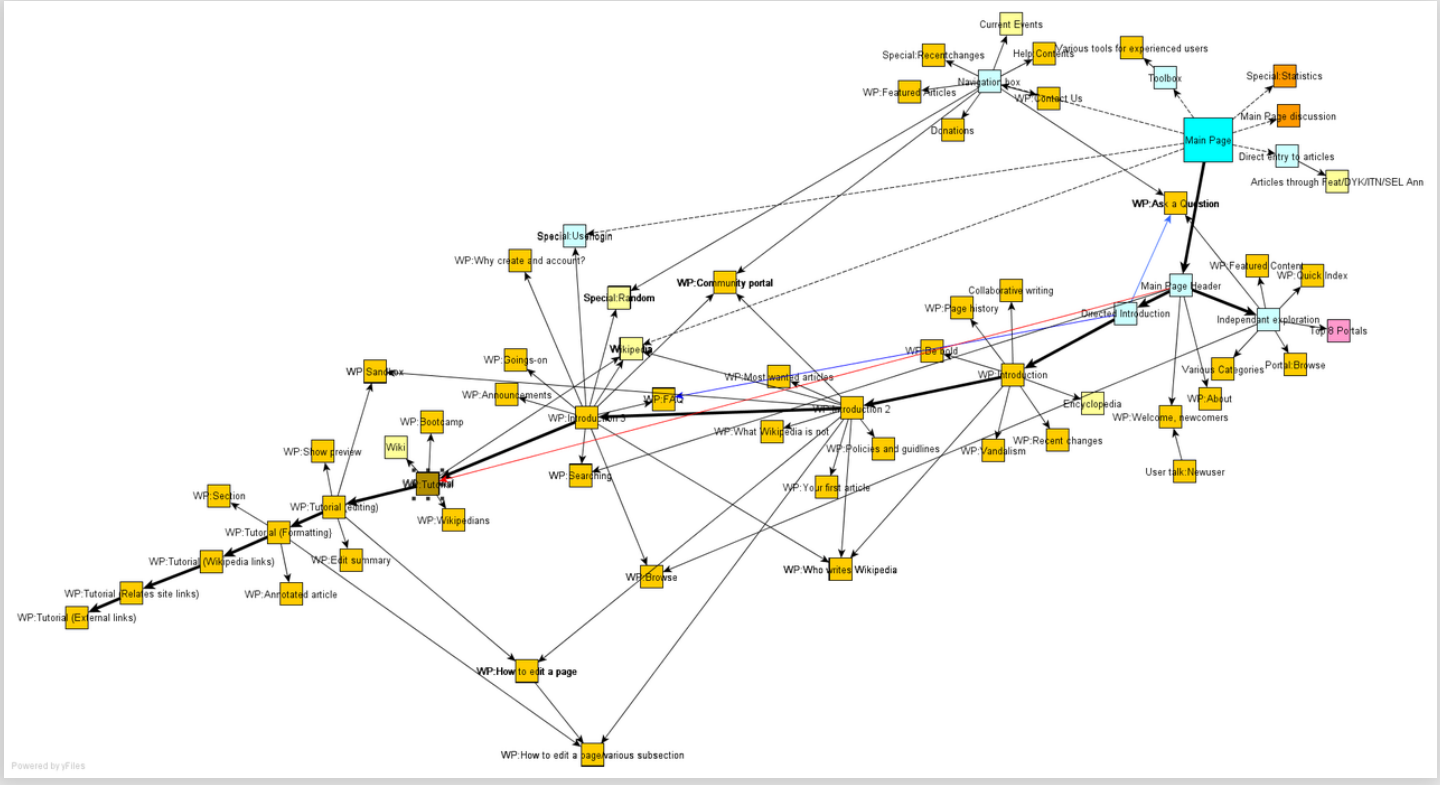
Many types of networks:

- Physical networks: telephone lines, roads, airline routes, rivers
- Information networks: WWW, citation networks
- Social networks: Facebook, Twitter, but not just: this class, Marriage between Royal Houses
- Biological networks: "food webs" (what eats what?), metabolical networks

Physical Networks

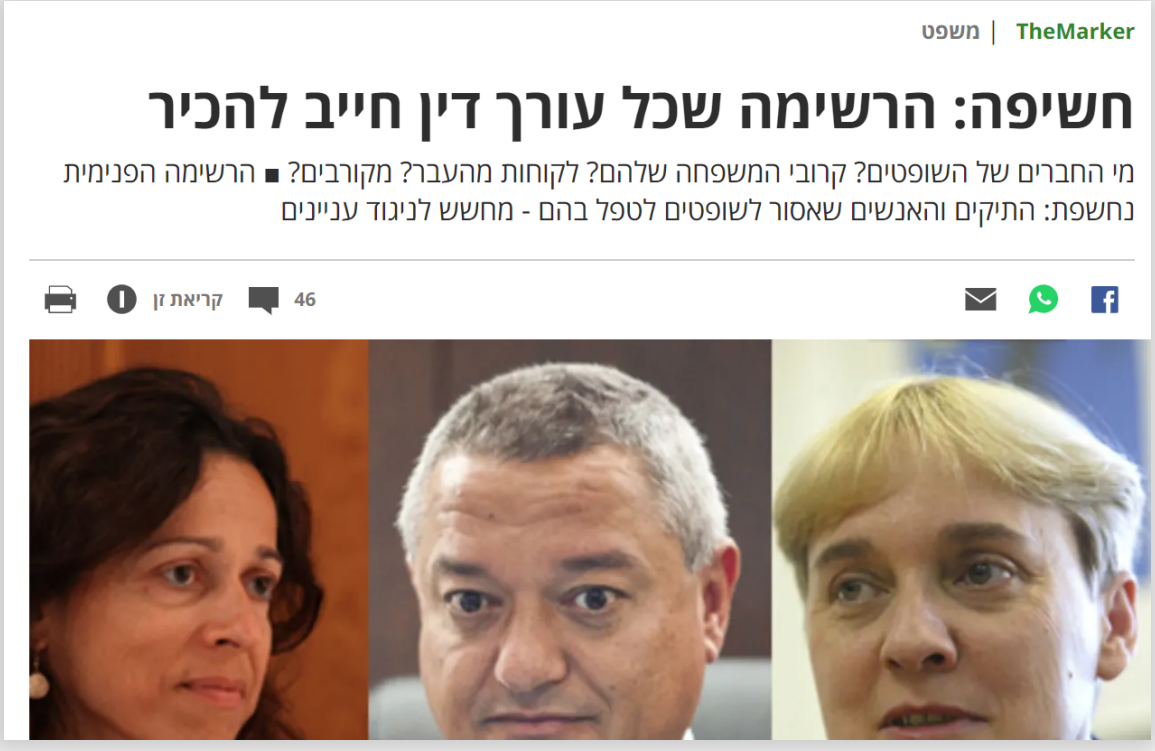
[מערכת להסעת המונים במטרופולין תל אביב]

Information Networks



[Wikipedia Main Page Site Map]

Social Networks



[חשיפה: הרשימה שכל עורך דין חייב להכיר]

Biological Networks

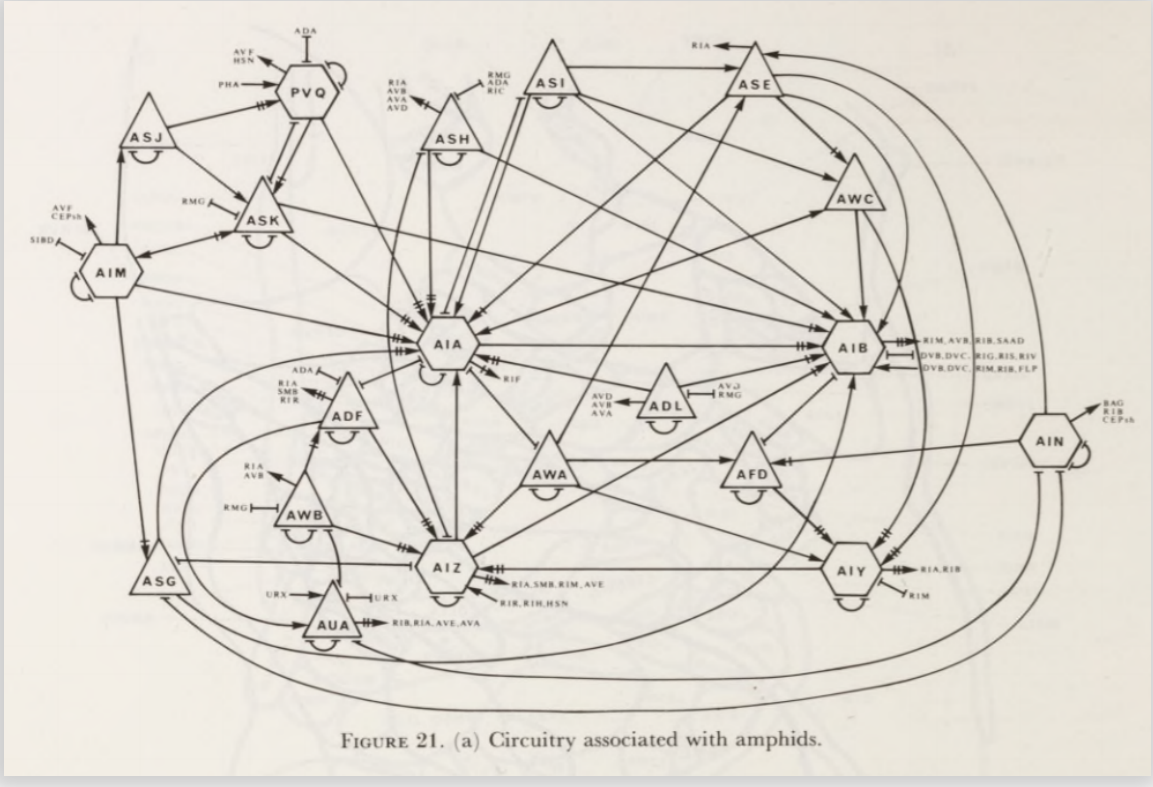
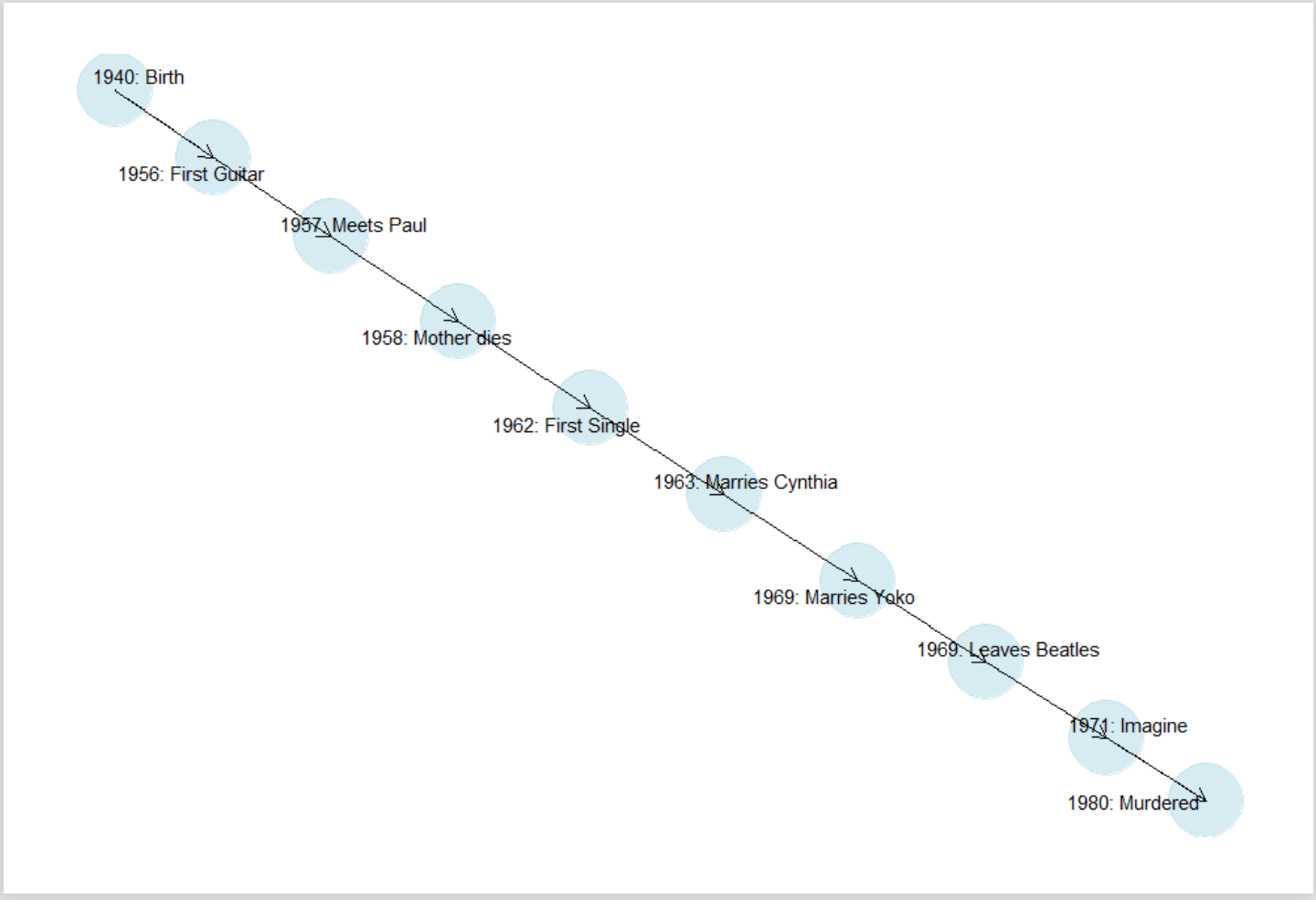


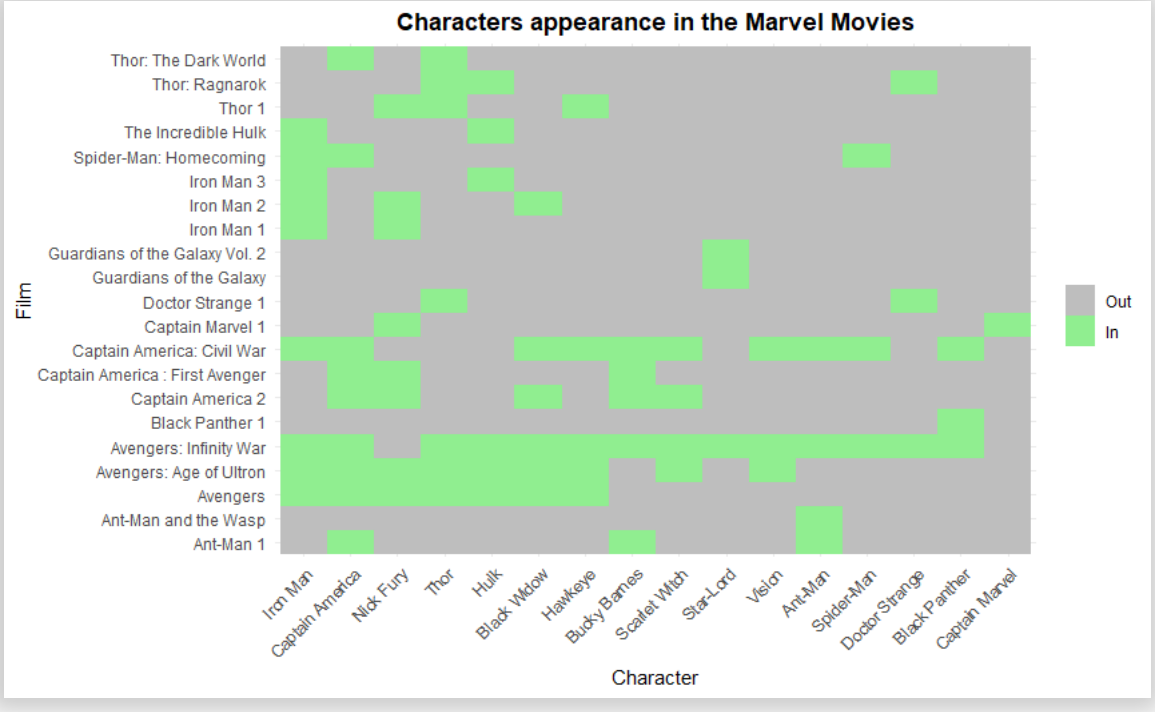
FIGURE 21. (a) Circuitry associated with amphids.

[The structure of the nervous system of the nematode Caenorhabditis elegans]

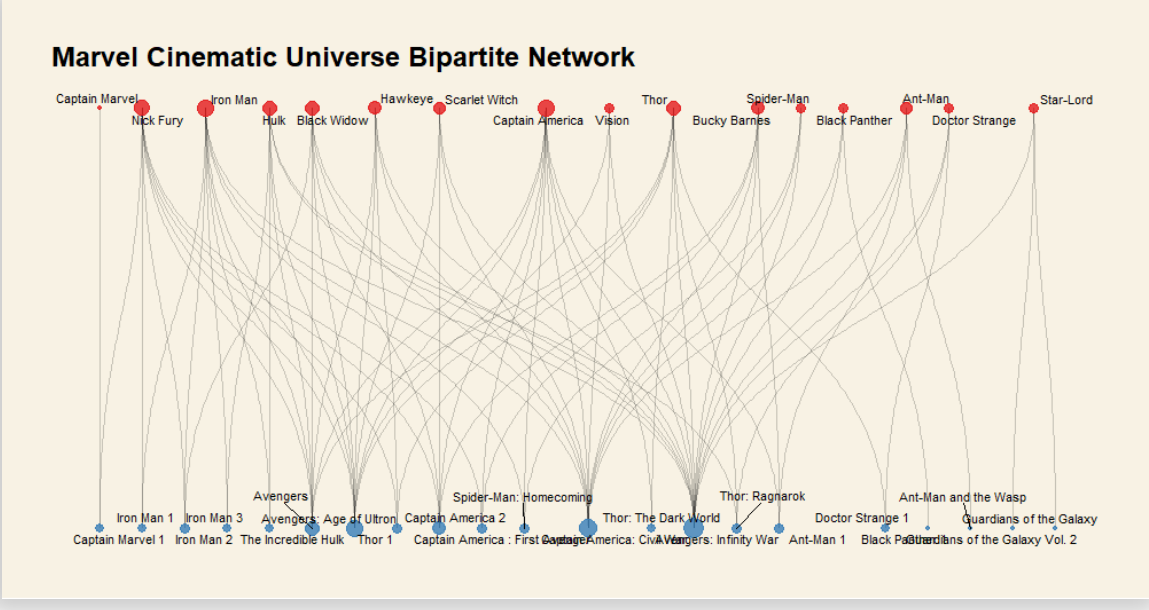
Harder to classify Networks



Bipartite Networks



Bipartite Networks

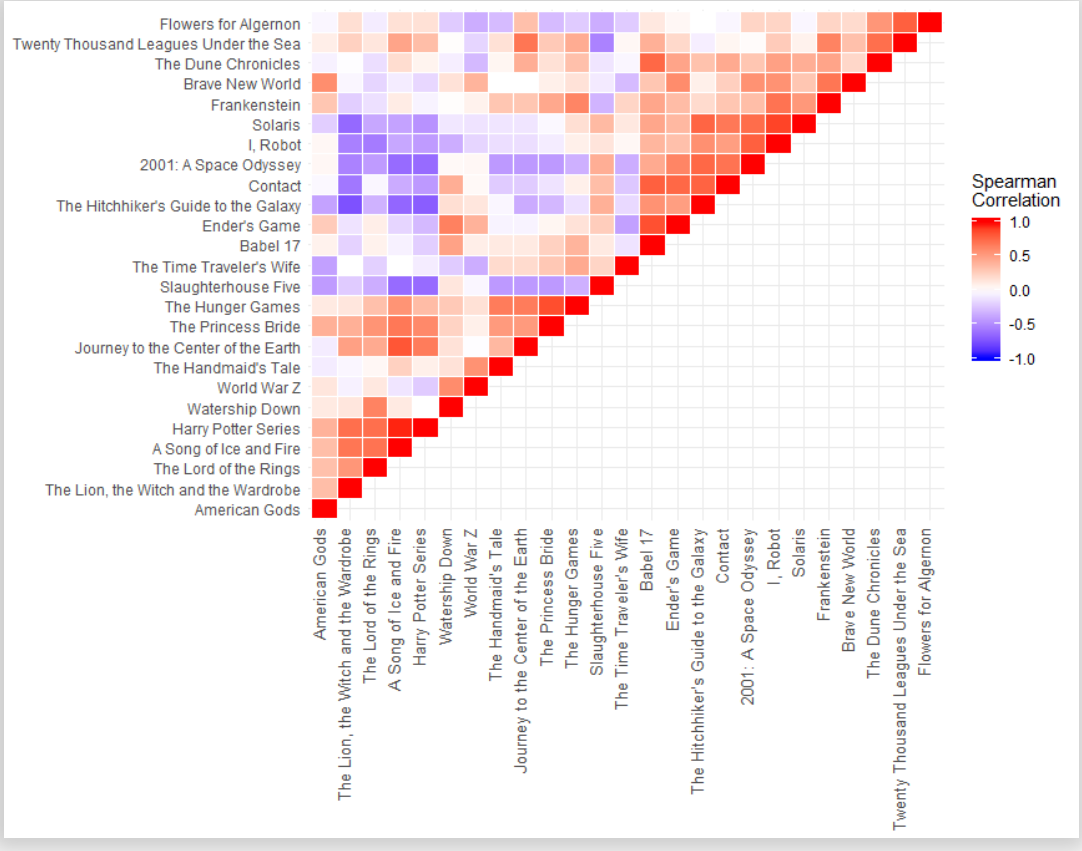


[Adapted from: Which Marvel Characters and Movies are the Most Central? / Félix Luginbühl]

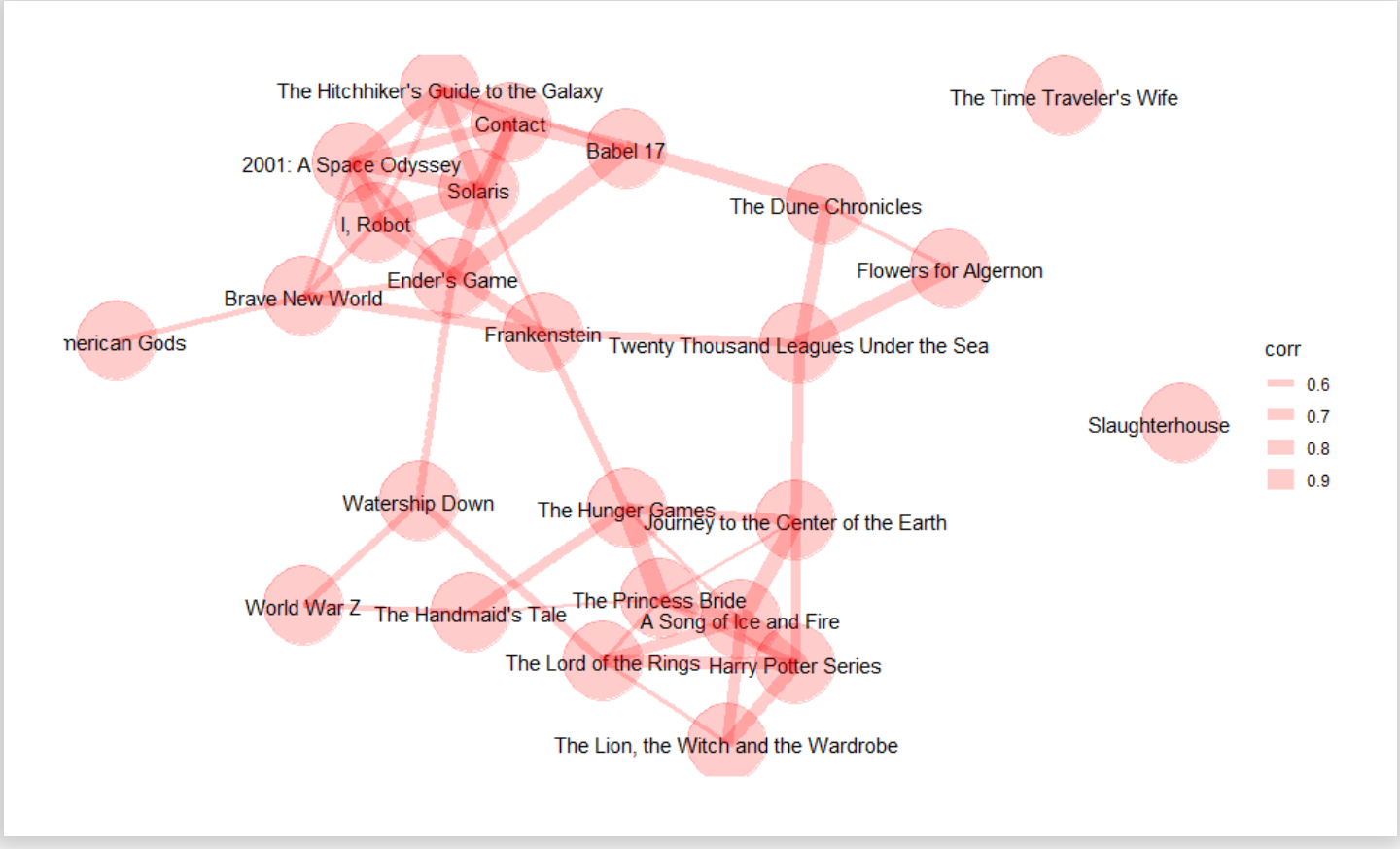
Similarity Networks

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	book	frequency	date	author	century	quarter_c	author_ge	robots_an	battles	romance	magic	time_trav	interplane	multi_spe
2	Frankenstein	6	1818	Mary Shel	1	1	2	1	1	3	0	0	0	0
3	Twenty Thousand Leagues Unde	7	1870	Jules Vern	1	3	1	0	1	1	0	0	0	0
4	Journey to the Center of the Eart	2	1874	Jules Vern	1	3	1	0	2	2	0	0	0	0
5	The Time Machine	6	1895	H. G. Well	1	4	1	0	2	3	0	3	0	0
6	The War of the Worlds	3	1898	H. G. Well	1	4	1	0	3	1	0	0	2	3
7	Princess of Mars	2	1917	Edgar Rice	2	5	1	0	3	3	2	0	2	3
8	The Skylark of Space	2	1919	E E Doc Sn	2	5	1	0	1	2	0	0	3	3
9	Last and First Men	3	1930	Olaf Stapl	2	6	1	2	2	0	0	0	0	3
10	Brave New World	6	1932	Aldous Hu	2	6	1	2	1	0	0	0	0	0
11	Swastika Night	2	1937	Katharine	2	6	2	0	0	0	0	0	0	0
12	The Space Trilogy	2	1938	C. S. Lewis	2	6	1	0	1	0	0	0	2	2
13	The Foundation Trilogy	8	1942	Isaac Asim	2	6	1	3	2	1	0	0	3	0
14	Chronicles of the Lensmen	2	1948	E E Doc Sn	2	6	1	1	3	0	0	0	3	2

Similarity Networks

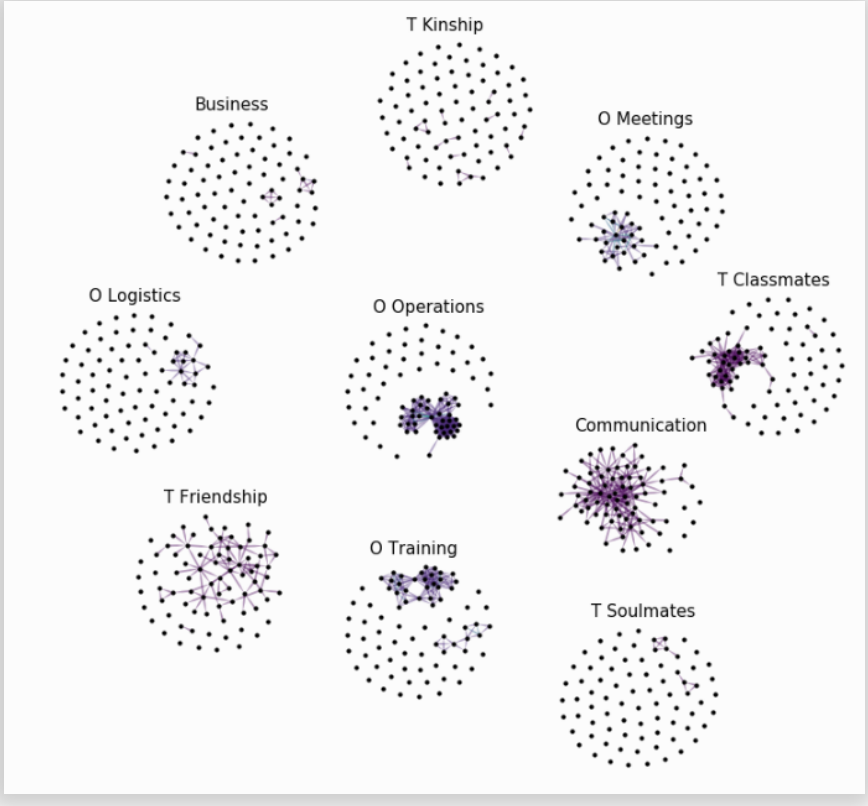


Similarity Networks



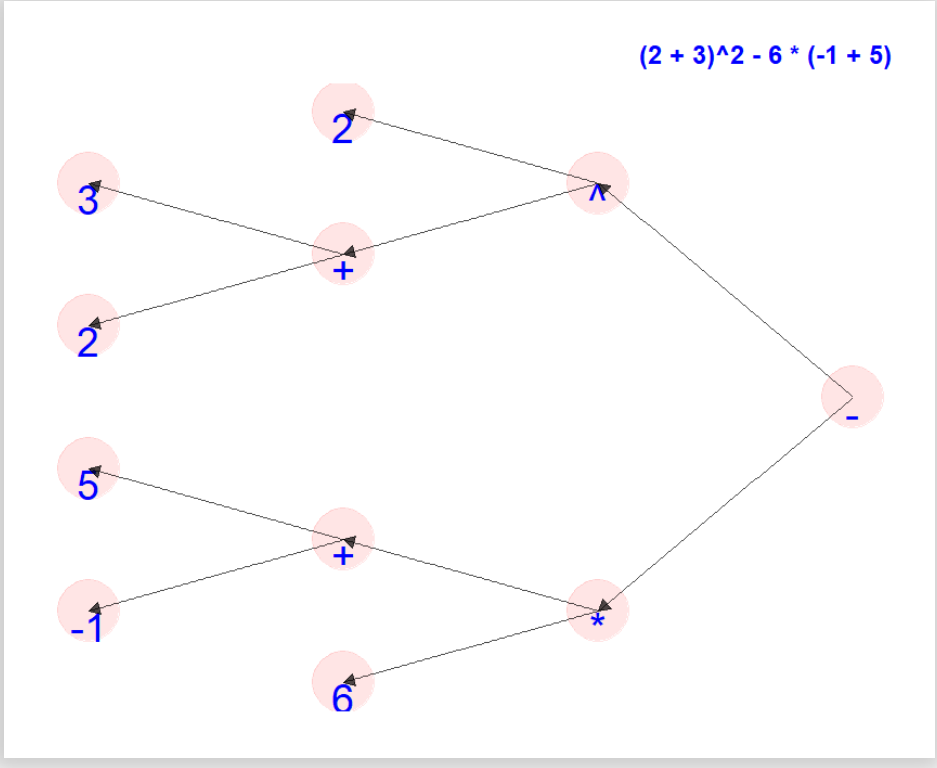
[Sci-Fi Books data set / Kathleen M. Carley]

Multilayer Networks



[Strategies for Combating Dark Networks (a.k.a Noordin Mohammad Top's terrorist network of South East Asia]

Trees



Does this remind you of anything?

Networks Properties

- Directed/Not: Do the edges have orientation?
- Weighted edges/Not: Do the edges have weight/strength/length?
- Connected/Not: Can you "get to" any node from any node?
- Self-edges/Not: Can a node be linked to itself?
- Acyclic/Not: Are there cycles? Can you get from one node to itself not by a self-loop?
- Time-varied/Not: Does the network change over time?

The networks we've seen

Network	Directed	Weighted	Connected	Self-edges	Acyclic	Time-Varied
Israeli Artists Coops						
Gush Dan Trains						
Wikipedia Site Map						
Israeli Judges Connections						
Worm Nervous System						
John Lennon Timeline						
Marvel Cinematic Universe						
Sci-Fi Books						
Noordin Terrorist Net						
Math Expression						

The networks we've seen

Network	Directed	Weighted	Connected	Self-edges	Acyclic	Time-Varied
Israeli Artists Coops			✓			✓
Gush Dan Trains			✓			
Wikipedia Site Map	✓		weakly		✓	
Israeli Judges Connections						✓
Worm Nervous System	✓	✓	weakly	✓		
John Lennon Timeline	✓		weakly		✓	👩
Marvel Cinematic Universe			✓			✓
Sci-Fi Books		✓				
Noordin Terrorist Net						
Math Expression	✓		weakly		✓	

Typical Questions About Networks

Macro:

- Is the network connected?
- Is the network dense or sparse?
- What is the maximum/average shortest path? small world effect
- Is the network homophilic for attribute X?
- Are there interesting communities in the network?
- Will a "message" percolate through the whole network? How fast?
- Can we model the network to give insight on how it developed? Predict how it *will* develop?

Typical Questions About Networks

Micro:

- Which is the "best connected" node?
- Which is the "most important" node? Not necessarily the same thing
- Is there a node or edge which "break" the network?
- Is there a path between node A and node B?
- What is the shortest path between node A and node B?
- Are nodes A and B likely to connect?
- What node should we recommend connect with node A?
- Can we predict missing attribute "X" for node A?

Obtaining Networks

- Surveys, Interviews, Questionnaires, Observations (Moreno's schoolchildren)
- Archives, sometimes historical (Padget's families of Florence)
- Snowball Sampling (Drug users' ego networks)
- Web Scraping (Adamic's political blogs)
- Web APIs (Twitter)
- Co-occurrence matrices (Marvel's cinematic universe)
- Any tabular dataset? (Sci-Fi books)
- Just really hard work (Milgram's Small world experiment)

The Adjacency Matrix

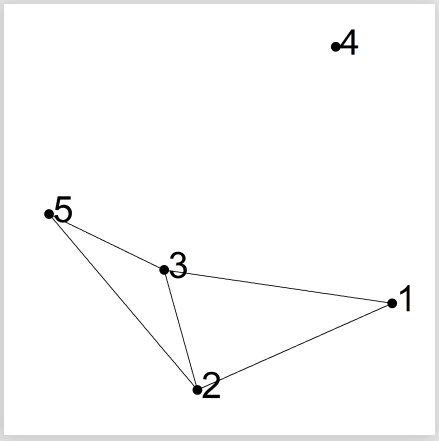
Undirected Networks

The Adjacency matrix A of an unweighted network is defined as a $n \times n$ matrix with elements A_{ij} such that:

$$A_{i,j} = \begin{cases} 1 & \text{if there is an edge between nodes } i \text{ and } j \\ 0 & \text{otherwise} \end{cases}$$

- The adjacency of a simple undirected network would be symmetric and contain only zeros on its diagonal
- If the network has multiedges - if there are q edges between elements i and j - then $A_{ij} = q$
- If the network has self-edges - if there is an edge between element i and itself - then $A_{ii} = 2$ (useful convention)

So this unweighted, undirected network:

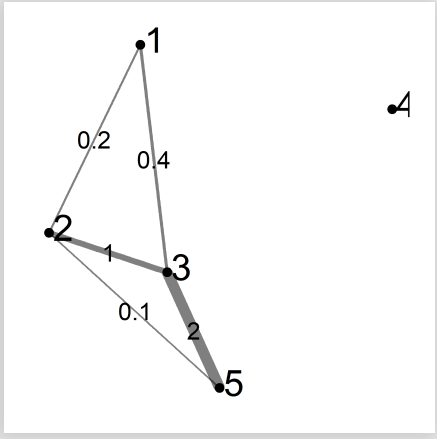


Would be represented as $A_{5\times 5}$:

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

For an undirected *weighted* network, A_{ij} would contain the weight.

- The higher the weight - the stronger the connection
- (In absolute value - because a weight needs not be positive)



$$\begin{bmatrix} 0 & 0.2 & 0.4 & 0 & 0 \\ 0.2 & 0 & 1 & 0 & 0.1 \\ 0.4 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 2 & 0 & 0 \end{bmatrix}$$

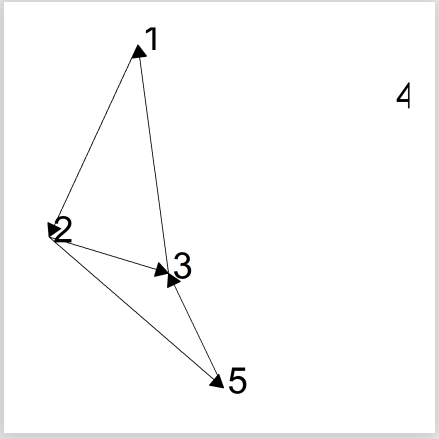
Directed Networks (a.k.a DiGraph)

The Adjacency matrix A of an unweighted network is defined as a $n \times n$ matrix with elements A_{ij} such that:

$$A_{i,j} = \begin{cases} 1 & \text{if there is an edge from node } j \text{ to node } i \\ 0 & \text{otherwise} \end{cases}$$

Note the convention from *column* index to *row* index can be confusing.

So this unweighted, directed network:



Would be represented as $A_{5\times5}$:

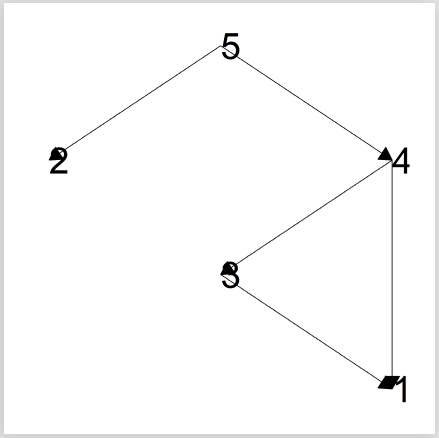
$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Directed Acyclic Graphs (DAG)

- If the network is directed and has no cycles, it is possible to draw the network such that all edges point downward, from a higher-numbered node into a lower-numbered node.
- If the network can be drawn where any edge from j to i implies that $j > i$, its adjacency matrix is *upper triangular*
- Since the diagonal contains only zeros (why) it is *strictly traingular*
- The other direction also holds (iff)

Is the netwok from previous page acyclic? Try both directions.

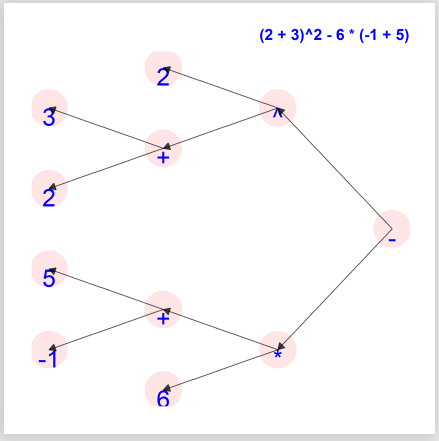
For example, this acyclic network can be numbered so that its adjacency matrix is strictly triangular:

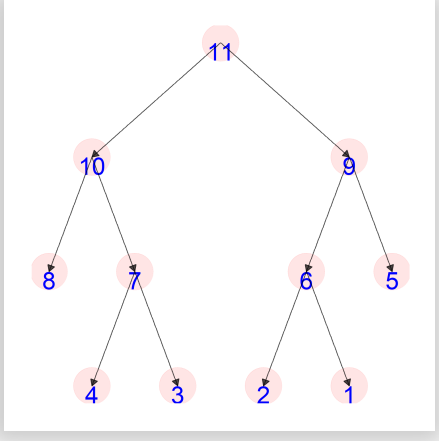


Would be represented as $A_{5\times 5}$:

$$\begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Example: the Math Expression Parse Tree





$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Bipartite Networks

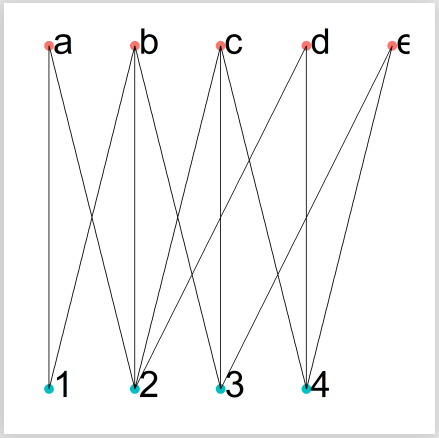
Incidence Matrix

We can write an Adjacency Matrix for a Bipartite Network (a.k.a *Two-Mode Network*) but a more compact representation exists: the Incidence Matrix.

If n items belong to g groups, the incidence matrix representing this bipartite network is $B_{g \times n}$ with elements B_{ij} such that:

$$B_{i,j} = \begin{cases} 1 & \text{if item } j \text{ belongs to group } i \\ 0 & \text{otherwise} \end{cases}$$

So this unweighted, undirected bipartite network of 5 items belonging to 4 groups:



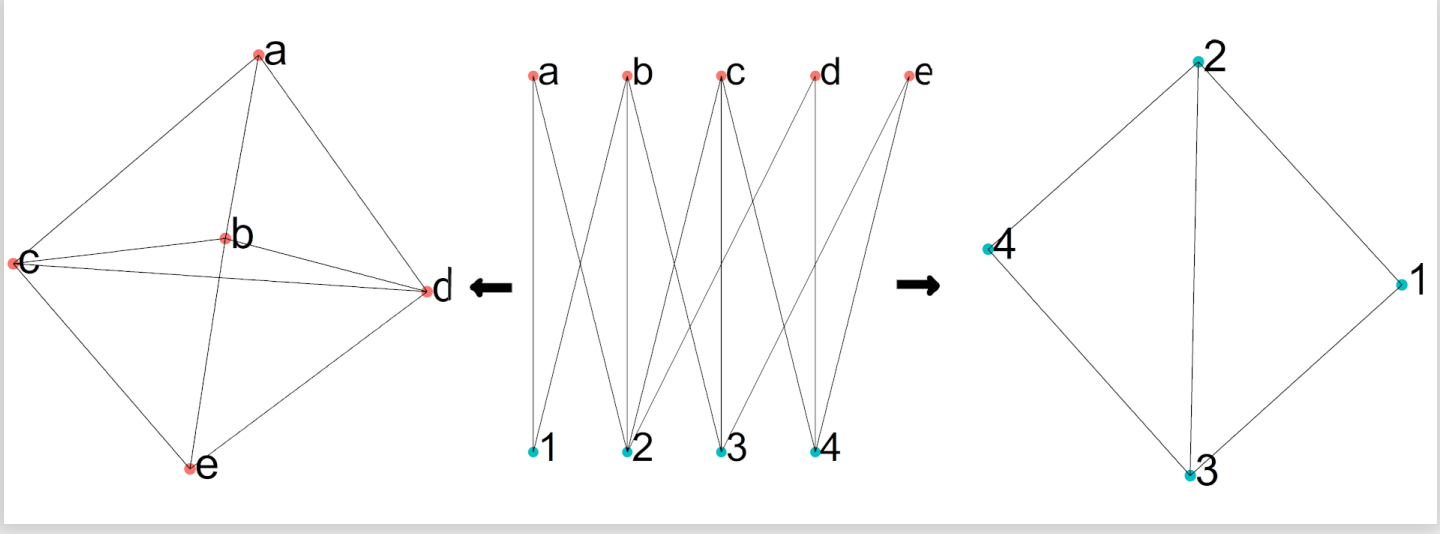
Would be represented as $B_{4\times5}$:

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Bipartite Networks Projections

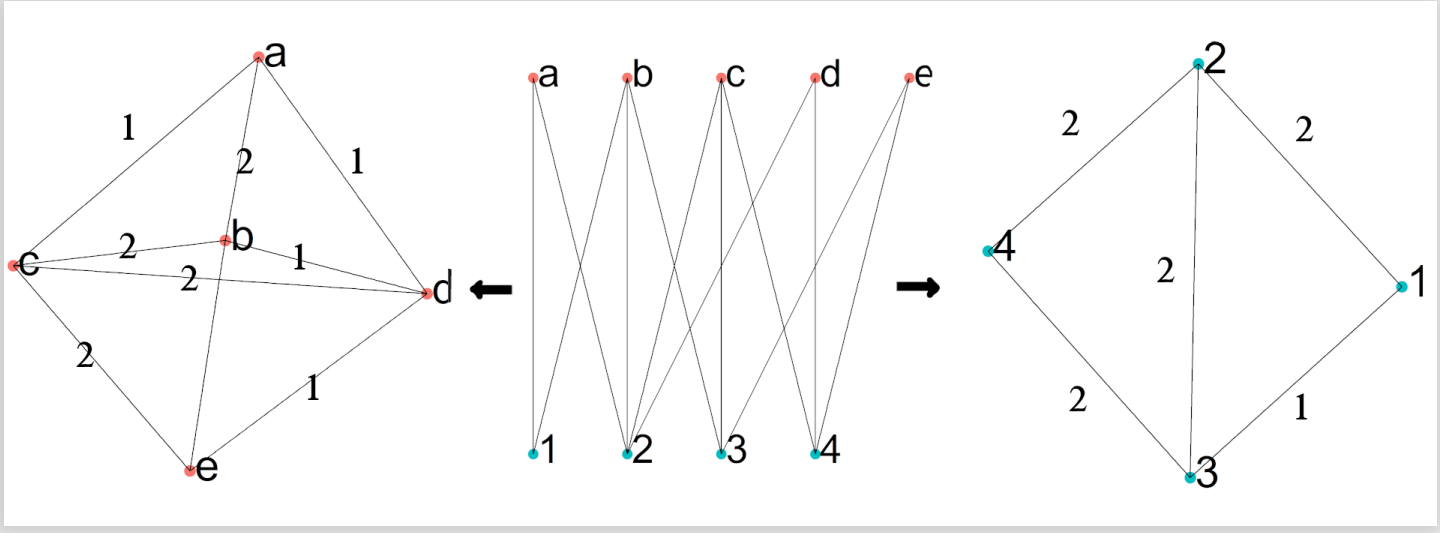
We can manually "project" a two-mode g groups x n items network into a one-mode undirected unweighted network of only g groups or only n items:

- If two items share a group - draw an edge between them.
- If two groups share an item - draw an edge between them.



However, this representation results in a loss of information.

For example, we could add for each pair of items (groups) how many groups (items) they share, with a weighted edge:



What does it mean in terms of incidence matrix B?

Exploiting the fact that B contains only zeros and ones, the number of groups shared by two items i and j, i.e. their *weight*:

$$w(i,j) = \sum_{k=1}^{g} B_{ki} B_{kj} = \sum_{k=1}^{g} B_{ik}^{\intercal} B_{kj}$$

And if we treat $w(i,j)$ as elements of matrix $P_{n \times n}$ we could write:

$$P = B^{\intercal} B$$

And P is *almost* the adjacency matrix of the items graph we've manually built. On its diagonal for each item i there is the total number of groups it belongs to:

$$P_{ii} = w(i,i) = \sum_{k=1}^{g} B_{ki} B_{ki} = \sum_{k=1}^{g} B_{ki}^2 = \sum_{k=1}^{g} B_{ki}$$

And so, if we wanted a proper weighted adjacency matrix, we would need to put zeros on P's diagonal.

How would we reach the $P'_{g \times g}$ adjacency matrix of the groups network?

Example: The Marvel Cinematic Universe Bipartite Network

```
import pandas as pd
import numpy as np

marvel = pd.read_csv("../data/marvel_incidence_matrix.csv")

marvel.shape
```

```
## (21, 17)
```

```
marvel.iloc[:4, :5]
```

```
##                     Film  Iron Man  Captain America  Nick Fury  Thor
## 0             Iron Man 1         1                0          1     0
## 1  The Incredible Hulk          1                0          0     0
## 2             Iron Man 2         1                0          1     0
## 3                 Thor 1         0                0          1     1
```

Get B of Marvel Films x Characters:

```
B = marvel.iloc[:, 1:].values
B.shape
```

```
## (21, 16)
```

Get P of Marvel characters:

```
P = B.transpose() @ B
P.shape
```

```
## (16, 16)
```

```
P[:5, :5]
```

```
## array([[9, 5, 4, 3, 5],
##        [5, 9, 4, 4, 3],
##        [4, 4, 8, 3, 2],
##        [3, 4, 3, 7, 4],
##        [5, 3, 2, 4, 6]], dtype=int64)
```

Why `P[1, 1] = 9`? Because Iron Man appears in 9 films.

```
marvel['Iron Man'].sum()
```

```
## 9
```

Why `P[1, 2] = 5`? Because Iron Man and Captain America appear together in 5 films.

```
marvel['Film'][(marvel['Iron Man'] == 1) & (marvel['Captain Americ
```

```
## 5                          Avengers
## 10        Avengers: Age of Ultron
## 12     Captain America: Civil War
## 15         Spider-Man: Homecoming
## 18         Avengers: Infinity War
## Name: Film, dtype: object
```

Putting zero in the diagonal to make P an actual adjacency matrix:

```
np.fill_diagonal(P, 0)
```

Degree and Density

Undirected Networks

The degree of a node in an undirected unweighted network is the number of edges connected to it (NOT number of its neighbors!).

The weighted degree of a node in an undirected weighted network is the sum of edges weights connected to it.

In terms of adjacency matrix A:

$$deg(i) = k_i = \sum_{i=1}^{n} A_{ij}$$

> Does this definition work for networks with multiedges or self-edges?

For an unweighted network: if m is the number of edges, there are $2m$ ends of edges. This is also the sum of the nodes degrees:

$$2m = \sum_{i=1}^{n} k_i = \sum_{ij} A_{ij}$$

> How would you define m to make this definition "work" for a weighted network?

This means that the average degree c of an unweighted undirected network is:

$$c = \frac{2m}{n}$$

The *density* of a network is defined as the fraction of existing edges from potential edges:

$$\rho = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)}$$

Which means the density for large networks is roughly the ratio of average degree to network size:

$$\rho = \frac{c}{n-1} \propto \frac{c}{n}$$

- If we could prove somehow that as $n \to \infty$ $\rho \to 0$, we'd call such network *sparse*. Alternatively, if ρ remains non-zero, we'd call such a network *dense*
- We can write $c = \rho n$. For some networks the average degree does not change. This means that as n grows ρ shrinks at rate $1/n$, and the network is called *extremely sparse*

💡 Think of an example of a network for which the average degree should remain constant as n grows.

Directed Networks

When adding directions to edges, we talk about the *in-degree* and *out-degree* of a node, as the number of ingoing or outgoing edges connected to it:

$$k_i^{\text{in}} = \sum_{j=1}^{n} A_{ij} \quad k_j^{\text{out}} = \sum_{i=1}^{n} A_{ij}$$

The number of edges m equals the sum of in- or out-degrees:

$$m = \sum_{i=1}^{n} k_i^{\text{in}} = \sum_{j=1}^{n} k_j^{\text{out}} = \sum_{ij} A_{ij}$$

The average in-degree is the average out-degree:

$$c_{\text{in}} = \frac{1}{n} \sum_{i=1}^{n} k_i^{\text{in}} = \frac{1}{n} \sum_{j=1}^{n} k_j^{\text{out}} = c_{\text{out}}$$

This means that the average degree c of an unweighted directed network is:

$$c = \frac{m}{n}$$

Which means the definition of density ρ remains the same for directed networks **in terms of average degree**:

$$\rho = \frac{m}{n(n-1)} = \frac{c}{n-1}$$

Example: The Marvel Cinematic Universe Characters Network

Let's convert the weighted undirected adjacency matrix P into unweighted:

```
P[P > 0] = 1
P[:4, :4]
```

```
## array([[0, 1, 1, 1],
##        [1, 0, 1, 1],
##        [1, 1, 0, 1],
##        [1, 1, 1, 0]], dtype=int64)
```

Get the list of degrees:

```
k = P.sum(axis=0)
k
```

```
## array([14, 14, 10, 14, 14, 14, 14, 14, 14, 13, 14, 13, 13, 13, 13,  1],
##       dtype=int64)
```

No. of nodes n:

```
n = P.shape[0]
m = np.triu(P).sum()
print('n nodes: %d; m edges: %d' % (n, m))
```

```
## n nodes: 16; m edges: 101
```

Average degree:

```
k.mean()
```

```
## 12.625
```

See that the simple definition is equivalent:

```
2 * m / n
```

```
## 12.625
```

Density ρ:

```
k.mean() / (n - 1)
```

```
## 0.8416666666666667
```

We can see that most characters are connected to most characters (16 overall). One character, however, is connected to only 1 character:

```
characters = marvel.columns[1:]
characters[np.argmin(k)]
```

```
## 'Captain Marvel'
```

is connected with:

```
characters[np.argmax(P[np.argmin(k), :])]
```

```
## 'Nick Fury'
```

Walks and Paths

Unweighted, Directed and Undirected Networks

A *walk* in a network is a route from node A to node B along the edges.

A *path* is a walk which does not intersect itself.

Walks and paths are extremely important in algorithms answering questions regarding a network's structure and the flow of information it. Of particular importance is the *shortest path* between two nodes.

Shortest?

The *length* of a walk/path of an unweighted network is the number of edges traversed along the walk (NOT number of nodes!)

For example, exploiting the fact that A contains only zeros and ones, we can easily compute the number of all walks of length 2 between two nodes:

$$N_{ij}^{(2)} = \sum_{k=1}^{n} A_{ik} A_{kj} = [A^2]_{ij}$$

(the ij-th element of the "squared" matrix A^2)

Similarly the number of walks of length 3 between two nodes:

$$N_{ij}^{(3)} = \sum_{k,l=1}^{n} A_{ik} A_{kl} A_{kj} = [A^3]_{ij}$$

And in general, the number of walks of length r between two nodes:

$$N_{ij}^{(r)} = [A^r]_{ij}$$

Shortest Paths

The shortest path (a.k.a *geodesic path*) between two nodes is the path of minimum length between the two nodes.

The *shortest distance* (a.k.a *geodesic distance*) is the length of the shortest path, in other words the smallest value of r such that $[A^r]_{ij} > 0$.

> Why is it not called "the shortest walk"? Must it be unique?
>
> What is the shortest distance between nodes which are not connected?

The *diameter* of a network is the maximal shortest distance between any pair of nodes.

> What is the meaning of the diameter in networks we've seen?

Components

Undirected Networks

- Some parts of a network are disconnected from each other
- A *component* is a subset of nodes such that there exists at least one path between each pair of nodes in the subset
- No other node in the network can be added and preserve this property
- A singleton node is a single component
- A network in which every pair of nodes are connected, has a single component and is said to be *connected*
- The adjacency matrix of a disconnected network *can be* written in block diagonal form

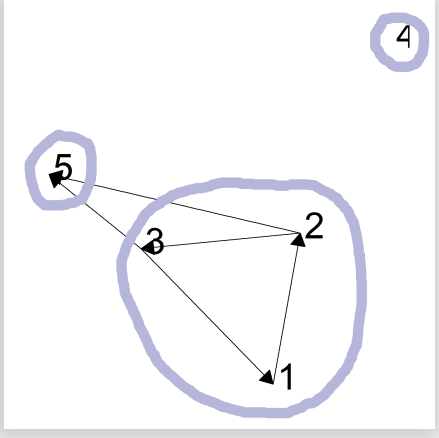
How many components are there in the Sci-Fi books network? Is it connected?

What is an easy way of making the network "more" or "less" connected?

Directed Networks

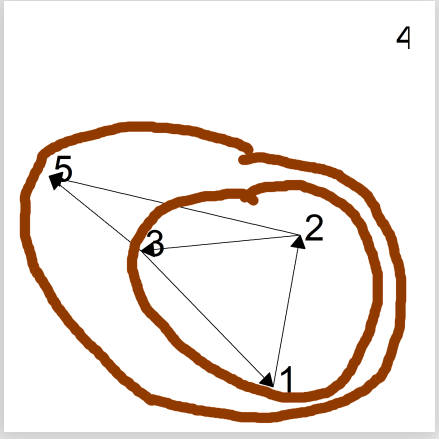
Type I Definition:

- *weakly connected* components: a subset of nodes such there exists at least one directed path between each pair of nodes in the subset, direction can be either way
- *strongly connected* components: a subset of nodes such that there exists at least one directed path between each pair of nodes, in both direction



Type II Definition:

- *out-component* of node i: subset of all nodes reacheable by a directed path from node i including i itself
- *in-component* of node i: subset of all nodes from which node i can be reached by a directed path, including i itself



The out-component of node 2 (or 1, or 3) is {1, 2, 3, 5}.

The in-component of node 2 (or 1, or 3) is {1, 2, 3}

The Graph Laplacian

Udirected, Weighted and Unweighted Networks

A different matrix representation of a network which proves useful in many situations is the Laplacian $L_{n \times n}$:

$$L_{i,j} = \begin{cases} k_i & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and there is an edge between nodes } i \text{ and } j \\ 0 & \text{otherwise} \end{cases}$$

Where k_i is node i's degree as defined before.

In other words:

$$L_{ij} = k_i \delta_{ij} - A_{ij}$$

Where δ_{ij} is the *Kronecker delta*, which is 1 if $i = j$ and 0 otherwise.

In other words:

$$L = D - A$$

Where A is defined as before, and D the diagonal matrix with nodes degrees along the diagonal.