Social Networks

Week 1

Code and resources : https://github.com/Schefflera-Arboricola/Stuff/tree/main/Projects%20and%20Simulation/Social%20Computing%20models

Some approximations ($n \to \infty$)

$^nC_2 \approx \frac{n^2}{2}$

$\log_{10} n \approx$ no. of digits in n

50 people → How many networks are possible b/w them? $\Rightarrow 2^{^{50}C_2} = 2^{1250}$

Way more than no. of atoms in universe

Main ideas

- Given a HUGE!! friendship network,

→ who is the most important person? Ranking nodes in a network (Page Rank)

→ how can we reach from one given node to other? path b/w 2 nodes or searching for a node

Small world phenomena → (*) Searching on a real-world network is suprizingly easy [for n nodes, searching $\equiv O((\log n)^2)$]

→ how can we predict weather 2 people are friends or not? "link prediction" (facebook's friend suggestions)

- Answering Philosophical Humanly Questions (are emotions and habits contagious?)
- Benefits of having acquaintances (in getting a job/spouse/new info.) (Strength of weak ties)
- How does information spreads in a network? How to make a viral video? How to market your product in a "social" network? (reach masses); How does fake news/believes spread in a "social" network?

Aditi Juneja

Week 2

Past :- Zachary Karate club (34 nodes, 78 edges)

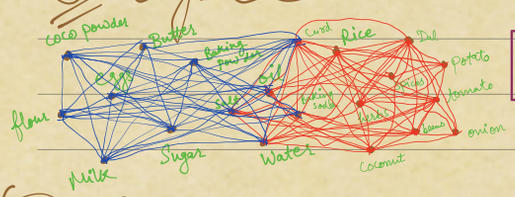
Now :- Huge data bcoz of internet

Objective :- Importance, analysis (crunching and inferences) of Social Networks Datasets

Playdoh (∞ shapes) ⟷ data (∞ inferences)

Examples of networks

1. Ingredients dataset (Network)



A — B means A and B are in same dish

→ Clusters : ingredients of a dish and also number of dishes

→ find the Most used ingredients (Ranking)

→ find common ingredients in 2 dishes

2. Synonyms Network

A — B means A and B are synonyms

love — like — same ... hate

→ This network would be connected. Why?

If it's connected, then there must be a path b/w opposite words

So, does this mean 'love' and 'hate' are synonyms? NOPE!

Degradation of Synonyms

Not a symmetric relation

A is synonymous to B and B is synonymous to C does not imply A is synonymous to C

3. Web Graph (Network)

(aka world wide web) (Page Rank)

A → B means webpage A have a hyperlink of webpage B.

→ We can find the best webpage, given a keyword

Hyperlink → links and also takes to the linked pag.

4. Social Networking sites' datasets

(facebook, linkedin, twitter, instagram)

Predicting if 2 nodes in the network are likely to connect, based on how the network is evolving, will increase connections of a person and hence their engagement on the platform. (link prediction)

Dataset formats

1. CSV (.csv or .txt ; cannot put attributes; comma separated edgelist or Adjacency matrix)

or space

```
1 2 0.3
3 4 0.1
5 7 0.2
4 7 0.8
```

```
1 2 5 7
2 4 6
3 1 4 7
4 6 2 3
5 7 2
6 1
7 2 8 4
8 9
9 1 7 4 8
```

2. GML (most common, we can put labels and attributes)

```
graph
[
  node
  [
   id A
   label "Node A"
  ]
  node
  [
   id B
   label "Node B"
  ]
  node
  [
   id C
   label "Node C"
  ]
  edge
  [
   source B
   target A
   label "Edge B to A"
  ]
  edge
  [
   source C
   target A
   label "Edge C to A"
  ]
]
```

```
graph
[ hierarchic  1
  directed  1
  node
  [ id  0
    graphics
    [ x 200.0
      y 0.0
    ]
  ]
  node
  [ id  1
    graphics
    [ x 425.0
      y 75.0
    ]
  ]
  edge
  [ source  1
    target  0
    LabelGraphics
    [
      text  "Happy New Year!"
    ]
  ]
]
```

3. Pajek Net (.net or .paj , flexible)

Keyword

# of vertices

edges

```
*Vertices 82670
1 "entity"
2 "thing"
3 "anything"
4 "something"
5 "nothing"
6 "whole"
*arcs
4244 107
4244 238
4244 4292 5
4247 107
4248 1
4248 54
```

4244 → 107

wt. of edge

Don't write if No labeling is needed

label of edge

this is below Node list

4. Graph ML (ML stands for XML; (.graphml); Kind of like HTML (tags), provides hierarchial structure, have attributes)

```
<?xml version="1.0" encoding="UTF-8"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns" xmlns:xsi="http://www.w3.org/
xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns http://graphml.graphdrawi
<key id="d0" for="node" attr.name="color" attr.type="string">
<default>yellow</default>
</key>
<key id="d1" for="edge" attr.name="weight" attr.type="double"/>
<graph id="G" edgedefault="undirected">
<node id="n0">
<data key="d0">green</data>
</node>
<node id="n1"/>
<node id="n2">
<data key="d0">blue</data>
</node>
<node id="n3">
<data key="d0">red</data>
</node>
<edge id="e0" source="n0" target="n2">
<data key="d1">1.0</data>
</edge>
<edge id="e1" source="n0" target="n1">
<data key="d1">1.0</data>
</edge>
<edge id="e2" source="n1" target="n3">
<data key="d1">2.0</data>
```

→ XML tag

→ graph ml tag

Name space

→ Meta data

attributes

→ graph tag

assigning green colour to d0 node

node tag

edge tag

↳ edge wt.

5. GEXF (Graph exchange XML format; created by Gephi, inspired by XML, clearner than GraphML)

```
<?xml version="1.0" encoding="UTF-8"?>
<gexf xmlns="http://www.gexf.net/1.2draft" version="1.2">
    <meta lastmodifieddate="2009-03-20">
        <creator>Gexf.net</creator>
        <description>A hello world! file</description>
    </meta>
    <graph mode="static" defaultedgetype="directed">
        <nodes>
            <node id="0" label="Hello" />
            <node id="1" label="Word" />
        </nodes>
        <edges>
            <edge id="0" source="0" target="1" />
        </edges>
    </graph>
</gexf>
```

XML tag

GEFX tag

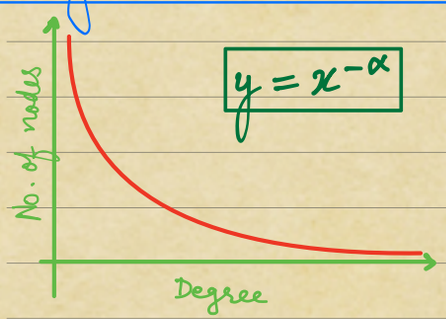
Meta tag

Graph tag

Nodes tag

edges tags

Degree distribution



$$y = x^{-\alpha}$$

(*) Most real world networks have power law degree distribution

↓

Very few nodes have very high degree and most of the nodes have comparatively low degree.

$$\boxed{\text{Density} = \frac{\text{Number of edges present}}{\text{Total possible edges}}}$$

↳ tells how closely knit the network is

nx.density(G)

complete graph = 1

no edge graph = 0

$$\boxed{\text{Clustering coefficient}(x) = \frac{\text{Number of edges among nodes directly connected to node } x}{\text{Total possible edges among nodes directly connected to } x}}$$

↳ tells how closely knit a particular node is in a network

↳ less for people who commit suicide. (i.e. people with lots of weak ties)

⊛ Avg(clustering coefficient(x)) = amt. of clustering present in a network

nx.clustering(G) → returns a dictionary of clustering coefficients

nx.average_clustering(G)

Diameter :- Shortest path b/w the 2 most distant nodes. nx.diameter(G)

⊛ In real world network the diameter is usually very small as compared to the number of nodes in the network.

Connectedness

V → no. of nodes ; e → no. of edges

$P(\text{after adding 1 edge, UAR, a particular node } y \text{ is isolated}) = 1 - P(y \text{ is not isolated})$

$$= 1 - \left(\frac{V-1}{{}^{V}C_2}\right) = 1 - \frac{2}{V}$$

$P(\text{after } e \text{ edges, } y \text{ is isolated}) = \left(1 - \frac{2}{V}\right)^{e}$

if $e = V$, $\left(1 - \frac{2}{V}\right)^{V} = \left(\left(1 - \frac{1}{V/2}\right)^{V/2}\right)^{2} \longrightarrow \frac{1}{e^2}$ (as $V \to \infty$)

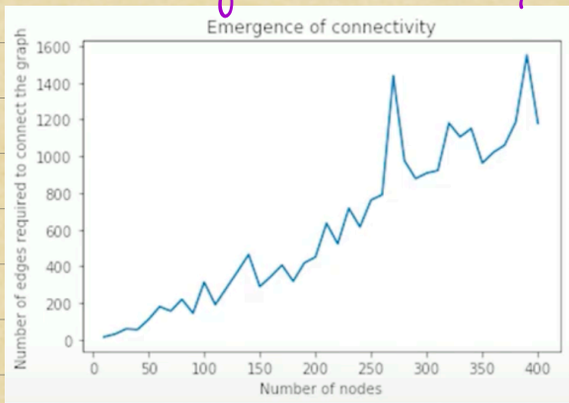
why nlogn? ⇒ Max. don diameter ques. extension.

if $e = V \log V$, $\left(1 - \frac{2}{V}\right)^{V \log V} = \frac{1}{e^{2 \log V}} = \frac{1}{V^2}$ ⇒ very small probability

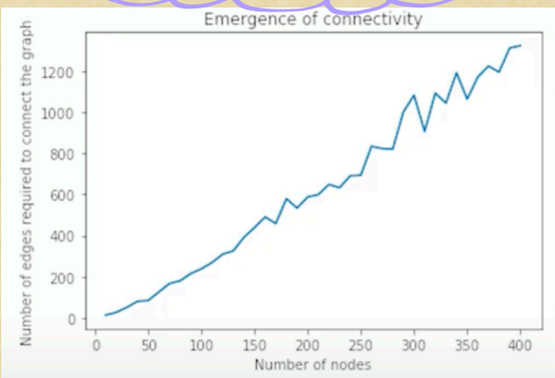
By putting nlogn edges b/w n nodes you will most probably get a connected network.

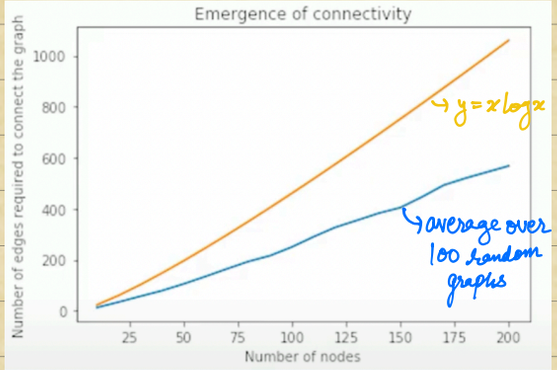
Analogy

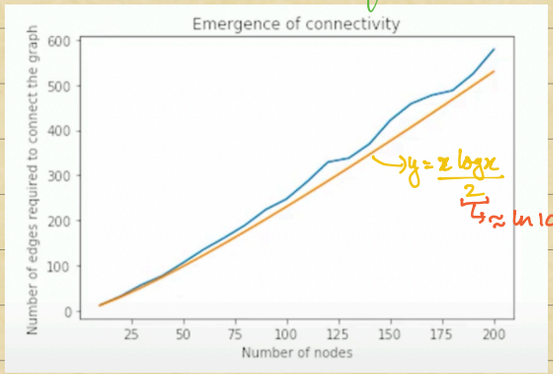
if we shoot a rectangle of area $\frac{n(n-1)}{2}$ UAR nlogn times it very unlikely that a big spot will be left unshot.



to get rid of spikes, instead of taking 1 random graph with n nodes, we will take 10 random graphs with n nodes and plot the average number of edges to join n nodes on y axis.







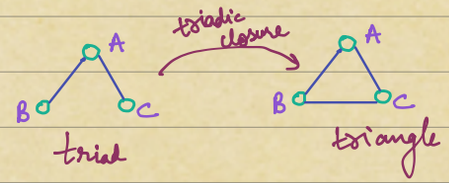
Aditi Juneja

Week 3 "Strength of weak ties"

Importance of acquaintance :- brings new information and prespective (breaks our social bubbles)

Keep your weak ties strong enough, but not so strong that they become your strong ties bcoz you cannot have more than a fixed number of strong ties, bcoz you are a human with limited time

Triads



A - B and A - C were strong ties.

BC can be a strong or weak tie.

Neighborhood Overlap

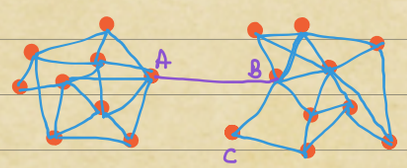
$$\boxed{\text{Neighbourhood overlap}(A,B) = \frac{\text{nodes connected to } (A \cap B)}{\text{nodes connected to } (A \cup B)}}$$

close to 1 means A - B is a strong tie

close to 0 means it's a weak tie.

$$\text{Richness} = \frac{\text{money spent wisely}}{\text{money earned}} \; ; \; \text{friendshipness}(A,B) = \frac{\text{common friends of A and B}}{\text{friends of A + friends of B - shared friends}}$$

Weak ties



$$P(\text{a random big graph would have 2 big disconnected components of sizes } n/2) = 1 - \frac{n^2/4}{n(n-1)/2} = \frac{1}{2(1-n)}$$

$n^2/4$ → possible No. of edges b/w 2 subgraphs

$n(n-1)/2$ → Total no. of edges possible

very close to 0 as $n \to \infty$

Why A-B might be a weak tie?

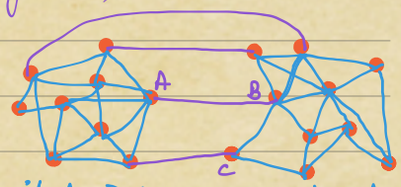
→ The above structure is highly improbable but possible

→ if A and B have a strong tie then A and C should also be connected bcoz of triadic closure

(*) edge A—B is called a 'bridge', removing A-B would break the graph. (Very Rare)

Local (weak) bridge : edge on which there is no triadic closure happening. (weak tie)

- → A's friends dolsnot know B and B's friends dolsnot know A. (true for bridge)
- → No common friends, but common friends are friends (Not true for bridge)
- → removing all local bridges will disconnect the graph.



here A—B is a local bridge

if A—B is a strong tie, then A-C must be friends too, but they are not, so A-B is a weak tie.

Granovetter's experiment :- most people got jobs by weak ties. But, he couldn't show local bridges were the weak ties, bcoz there was no way to quantify strength of a tie. But, in cell phone network, 'call time' can be used as quantification.

Cell phone network (almost connected)

$A \xrightarrow{c} B$ means A talked to B for c minutes.

call time ≡ strength of tie

low neighbourhood overlap ≈ local bridge

Result :- edges with low N. overlap had low call time.

$\boxed{\text{neighbourhood overlap} \propto \text{strength}}$

Embededness

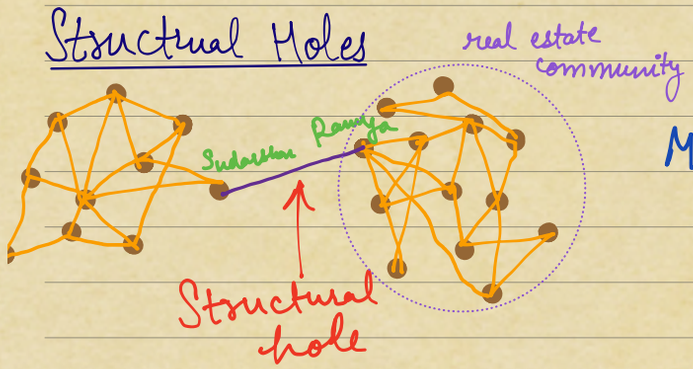
(common friends ↔ trust)

Should we have high embededness in our relationships?

→ yes :- More strong ties, more trust, high penality for breaking trust

→ no :- imp. of acquaintances, good for bussiness

Structual Holes



Ramya is a real estate agent.

Monopoly of Ramya :- everyone in Sudarshan's community will go through Ramya to know about real estate.

Social Capital (facebook's capital → user's attention)

So how should facebook shape the friendship network, so that the engagement maximizes?

1. Build a certain number of friends, by friend suggestions based on kind of posts one likes / sees.
2. These friends should post interesting (engaging) things

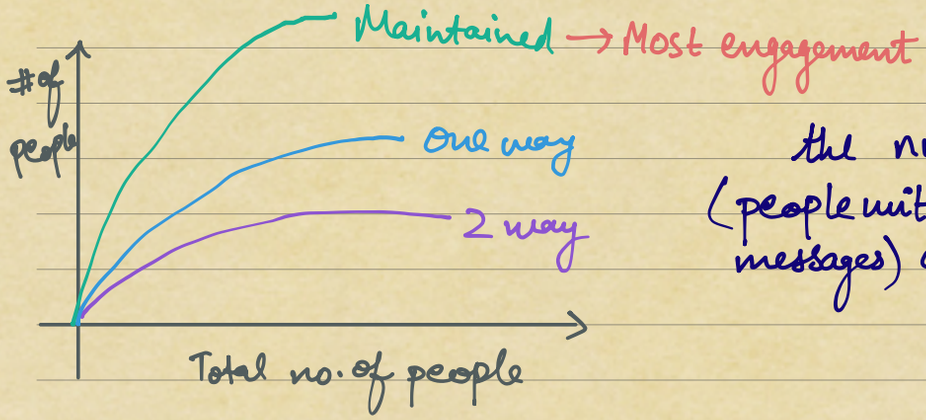
Best friendship network is a mix of closures (friend's friend is a friend) and brokerages (structural holes / bridges).

Types of engagements in a S.N.

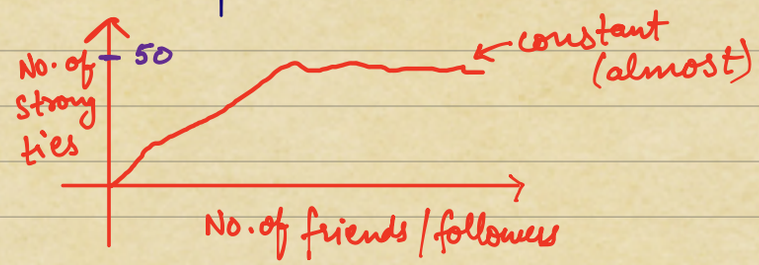
Passive engagement When one person is interested in other person's posts but the other person is not (Maintained relationships) Most common

One way communication negligible comm.

2 way or mutual comm. strong friendship. (aka reciprocal relation)

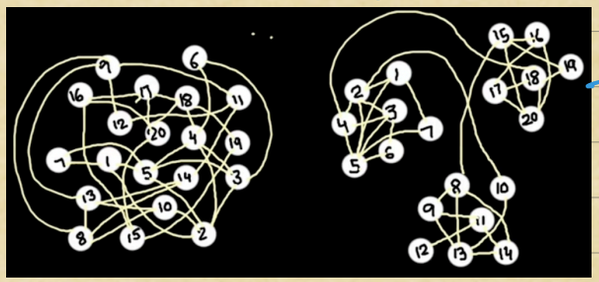


A person may have lots of friends but the number of strongly connected friends (people with whome the person had exchanged 5–6 messages) doesn't cross a particular threshold.



Community detection

We need to define what a 'community' is, bcoz we cannot see them always



→ we cannot rely on our eyes.

Community:- bunch of nodes where there are a lot of edges in b/w them but not across other parts of the network

(a little less vague definition)

Road map analogy ⇒ highways connect cities,

highways ↓ bridge; cities ↓ community

→ a lot of roads inside a city as compared to across cities.

detect communities ≡ remove these 'highways'

How to find these 'highways'? (Betweenness)

how much does an edge come in-between?
Sudarshan — Ramya edge comes in between a lot.

$$\text{Betweenness (edge)} = \frac{\text{no. of shortest paths from node 'a' to 'b' passing through this edge}}{\text{Total no. of shortest path b/w 'a' and 'b'}}$$

$$\boxed{\text{Betweenness}(e) = \sum_{\forall s,t \in V} \frac{\sigma_{st}(e)}{\sigma_{st}}}$$

node list

Aditi Juneja

Week 4 "evolutionary model"

Objective:- How do networks evolve?

Friendship network

→ how much does your friends influence you?

→ "Tell me who your friends are. I'll tell you who you are."

→ parent's advice: keep good company (your company is influenced by their companies and their companies by their companies' companies and so on!)

→ Christakis and fowler's study's conclusion:- obesity / happiness are contigious sometimes upto 3 levels.

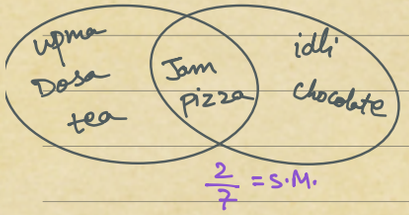
→ How do we pick friends? (do we pick 'our types' or converts randomly picked people to 'our type' or becomes 'their type'?)

Selection :- People with similar traits become friends. Eg:- french speaking people, eco. status, health habits

"Birds with the same feathers flock together."

Social influence :- When other people influence you and their traits starts becoming yours. Eg:- smoking, health habits.

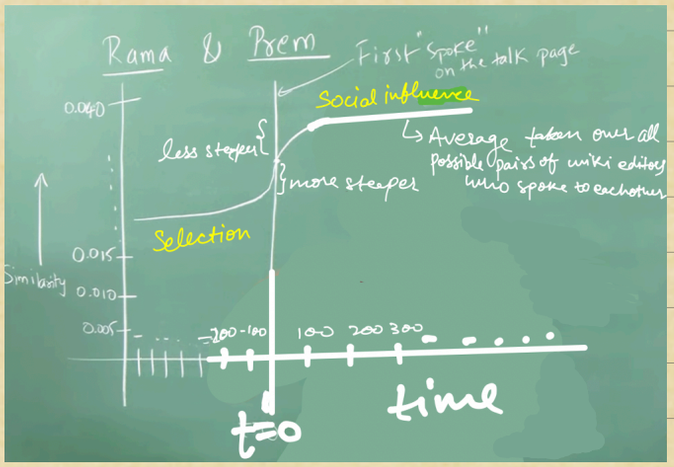
In a social network, what amount of friendships have social influence in them and what amount of them are selections?



$$\text{Similarity measure}(A, B) = \frac{A \cap B}{A \cup B}$$

A and B are set of traits of personA and person B.

for Wikipedia dataset :- S.M. = $\frac{\text{pages edited by person1 and person2}}{\text{Total no. of pages edited by both of them}}$



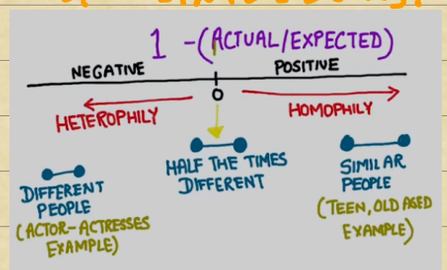
Conclusions

- → Sudden inc↑ in S.M. before 2 people speaks
- → a less sudden inc↑ in S.M. after 2 people speaks
- → overall people who talk to each other becomes more similar to each other than they were before.

HOMOPHILY (Grp1, Grp2) = $1 - \left[\frac{\text{Actual no. of edges b/w Grp1 and Grp2}}{\text{Expected no. of edges b/w Grp1 and Grp2}}\right]$

→ $\frac{E}{2}$ if $n(Grp1) = n(Grp2)$, E → # of edges

- → Measure of 'how exclusive grp1 and grp2 are'.
- → like attracts like
- → here, groups are a set of nodes with certain similar traits.



= 1 (No homophily, No hetrogeneity)

> 1 (Hetrogeneity) Grp1 and Grp2 are very inclusive of eachother

< 1 (homophily) Grp1 and Grp2 are very exclusive of eachother

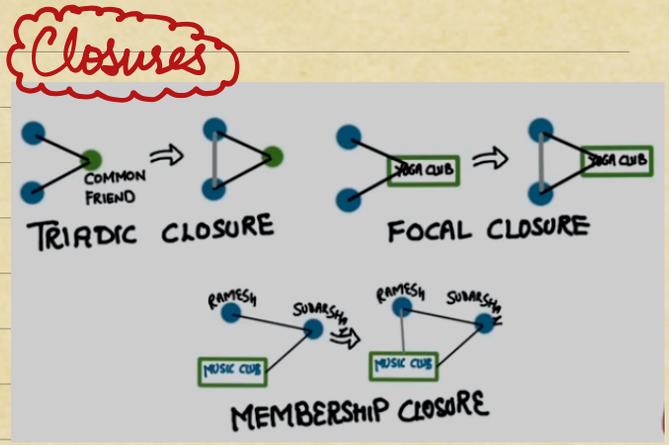
Difference b/w 'Homophily' and 'selection'?

selection leads to Homophily

Quantifying effects of triadic closure:

(obs. a friendship Network over time)

→ approx. Linear (actually exp.)



$$P(\text{A and B becoming friends with K common friends}) = \boxed{1-(1-p)^K}$$

[here, p is the independent probability that each K common friends contributes towards creating a link b/w A and B.]

P(2 people becoming friends) $\propto$ No. of common friends

Evolutionary model = Homophily + 3 closures + social influence

Aditi Juneja

Week 5 "Stabalising Networks"

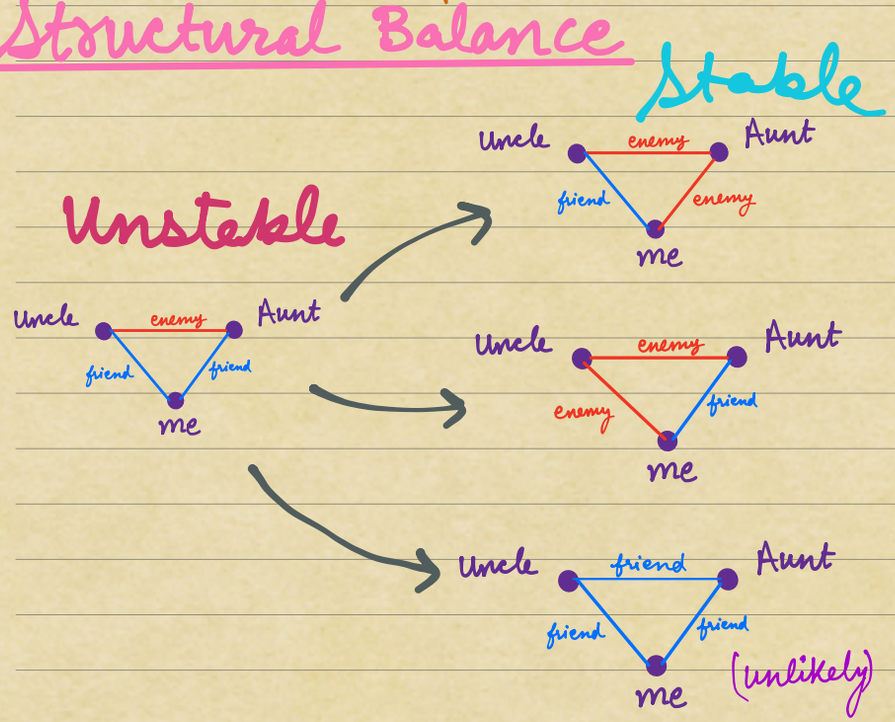
if you are within 'your type' of people mostly then probably you will not migrate

threshold = t %

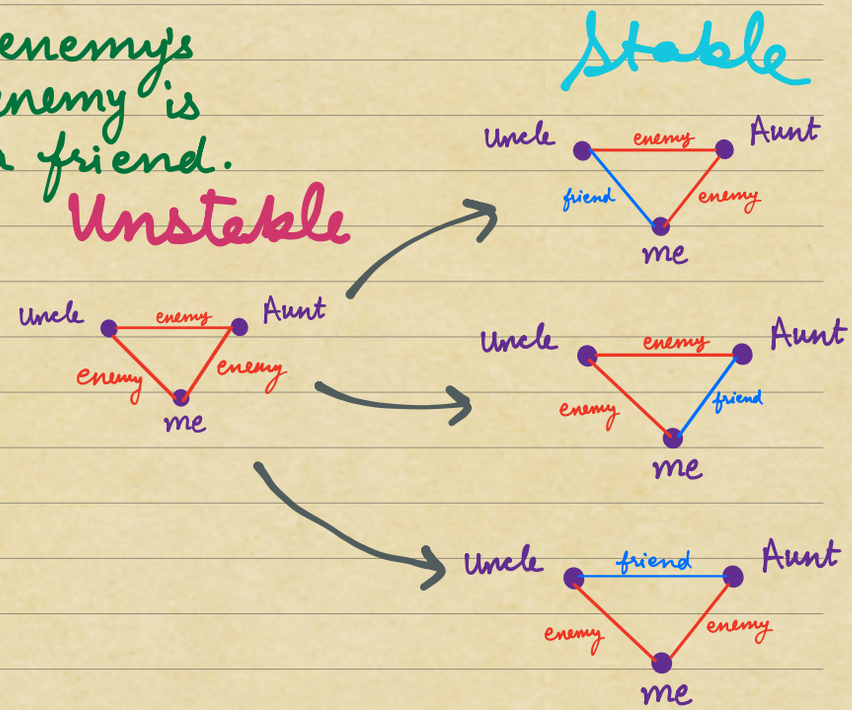
Schelling Model :- If we put random people in a box, then they tend to segregate.

☆ Amount of homophily / segregation depends on individual bias (i.e. t), empty space, ease of migration, population size.

Structural Balance



enemy's enemy is a friend.

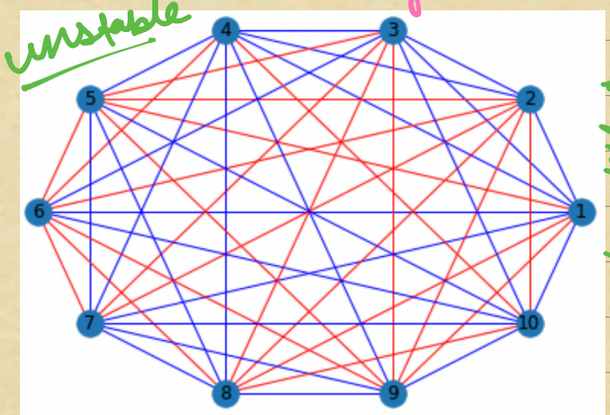


(*) Bonding increases due to similar likes and dislikes, when 2 people struggle together, have a common enemy / friend.

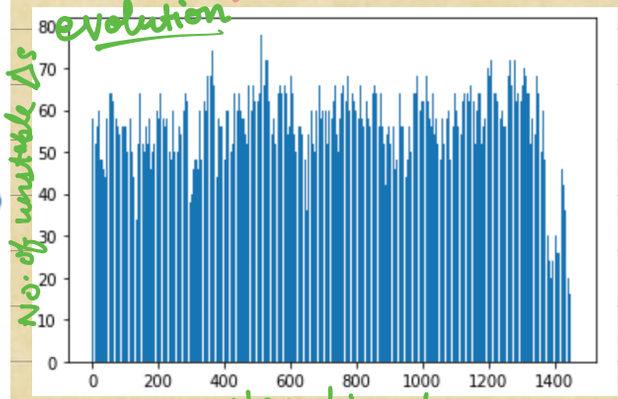
In a network, if 1 triangle gets stabilized then neighbouring triangles might become unstable. So, How will a stable graph look like?

Balance Theorem Given a random graph with positive and negative edges the balancing of graph will eventually form 2 groups with positive edges within and negative across. (Proof:- separately write all the +ve and -ve relations of 1 node and observe!!

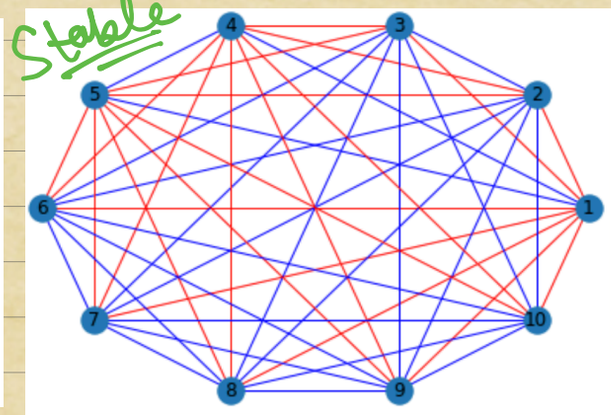
unstable



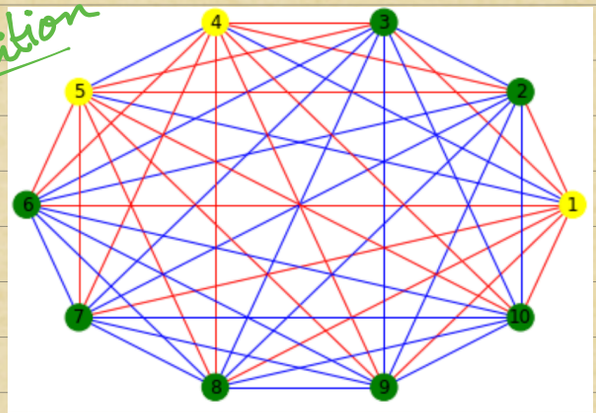
evolution

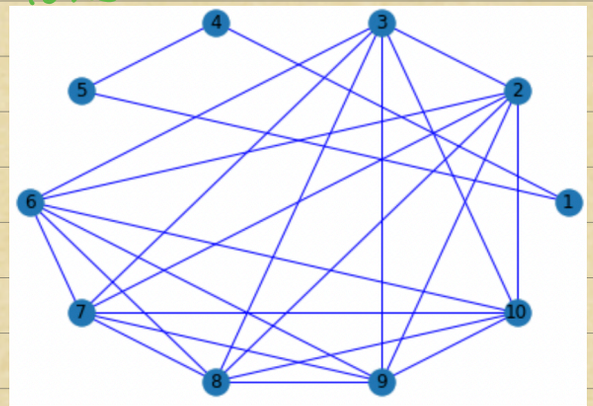


Stable

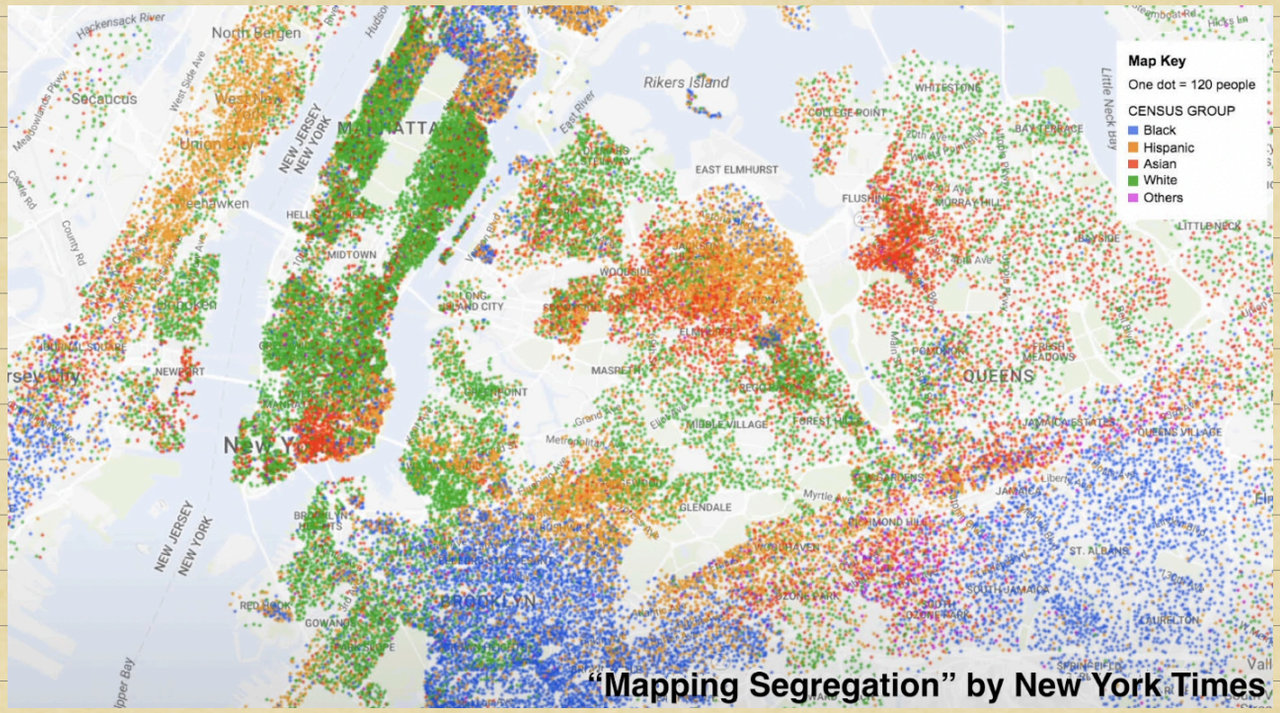


Coalition





removing all negative edges gives us 2 disconnected subgraphs. (or coalitions)



“Mapping Segregation” by New York Times

Aditi Juneja

Week 6 "Searching on web"

Objective :- Given a Keyword, how can we rank the webpages, in the web of hyperlinked webpages, from most to least appropriate?

Solutions

1.) Maintain database
   → not scalable
   → tedious

2.) Understand the web
   a.) collect the web network
   b) Use web graph to search

a.) To make the world map, you would have to randomly explore the world. So, to get the web network, walk randomly on the graph until you have explored enough
   ↳ what's that? ~ $n \log n$ steps (also, the no. of edges to connect the graph (week 2))

b.) Searching

① Equal coin distribution method :- Give 100 coins to all explored nodes, then redistribute coins, such that coins of a node are equally distributed among it's neighbours, until the distribution of coins becomes constant.

② **Random Walk** :- Pick a random node, give it a coin, go to any of it's neighbour, (if no neighbour, go to a random node), give it a coin, then go to it's neighbour and so on, until the proportion of coins with each node becomes constant. (Larry Page and Sergey Brin)

→ Both methods are fundamentally same.
→ Rank the nodes based on the no. of coins.

Who has the most coins? → the node whose neighbours give a lot of coins and that would mean neighbour's neighbours are rich and so on...
AND rich neighbours should not have alot of neighbours.

How can you ensure that your rich cousin keeps giving you money?
① ensure cousin keeps earning
② ensure she doesn't lend money to alot of people

Who are famous? → who famous people thinks are famous. (not related to indegree)
→ bcoz, 10 famous people thinking I am famous is different from any 10 people thinking the same

GNU → GNU Not Unix

Recursive definitions!!

How does google works? Given a Keyword, it collects all the webpages, from the explored web, having that Keyword and ranks them using random walk method (aka PageRank Algo.) [fun fact: 'page' in 'PageRank', refers to both ranking web 'pages' and Larry 'Page']

Aditi Juneja

Week 7 "Diffusion in network"

Do we follow each other? → Yes

What do we mean by 'follow'? We 'follow' a group or someone when we:-

1. start adopting their habits (Smoking, obesity)
   ↳ 3 level influence
2. buying things that they buy (ipad)
3. start getting some kind of info. from them. (journalists | friends)

Why do we follow? → conformity
→ Monetary Benefits (Pyramid Scheme example, farmer following other farmers benefits her yeild)
→ New Information (Doctor following other doctors to be aware of new medical advances.)

Should we follow? Should we take 'the road not taken'?
Is it good for us to follow, as individuals and as groups?

→ In a 'follow' network, how will the trajectory of a product / idea, that is diffusing through the network, look like?
→ How can we predict, how many people will adopt the product/idea, at the end?

Me → complete Assignment (Library) → 2 (payoff)

Me → enjoy (mall) → 3 ✓

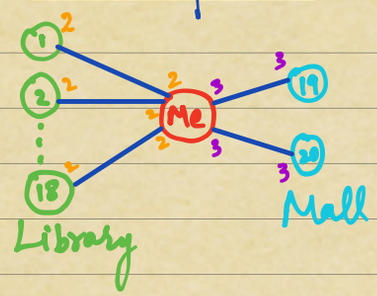
Now, 20 friends → 18 want library, 2 want mall (Assumption:- All friends equally influence you.)

analysis of 1 pair:

	Choice 1	Choice 2	Choice 3	Choice 4
Me	Mall	Library	Mall	Library
friend	Library	Mall	Mall	Library
payoff	0	0	3	2

✓

Whole picture:-

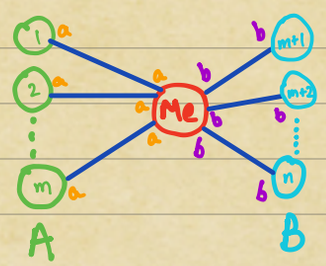


what to do now?

if library, payoff = 18 × 2 = 36 ✓

if mall, payoff = 2 × 3 = 6

Generalise :-



$p = m/n$

for A, $pa \geq (1-p)b \Rightarrow p \geq b/(b+a)$

for B, $pa \leq (1-p)b \Rightarrow p \leq b/(b+a)$

if more than p% of my friends

Social Reinforcement when you tend to belief something is true, because you hear it from a lot of people.

everyone must be weighing their options, just like me, so what will be the final seen?

initially :- 2 out of n wants to go to mall.

Case 1 :- 2 people change their mind and everyone studies.

Case 2 :- everyone flips and everyone goes to the mall.

Case 3 :- 2 communities get formed, all the people in community of 2 rebels, and also in the communities closely connected to them, will go to mall and rest to Library.

Quantifying 'closeness' or 'unity'

$$\text{density of a community} \propto \frac{1}{\text{ease with which an idea can enter it}}$$

This is how usually a class looks like.

Every node have a threshold p.

Complete cascade :- The idea would spread all over the network.

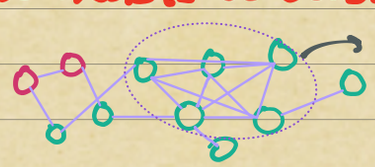
Cluster density :- If each node in the cluster (or subgraph) has at least a 'p' fraction of its neighbors in the cluster (or subgraph).

(avg. proportion of friends in cluster)

Condition of complete cascade :- p ⇒ threshold for doing a task, for each node

If there exists a cluster of density 1-p, then complete cascade is not possible.

p = 40%



for these nodes threshold won't pass and hence they will remain green

⊛ Another interesting thing:- Incomplete cascade $\Longleftrightarrow$ cluster of density $\geq 1-p$ exsists.

Problems in launching a new product :-

→ It's easier to do something that people around us do.

→ But, everyone is hesitant to try a new product. (time machine)

Solution → Advertise to people who are famous aka Viral marketing. (targetted ads. are even better)

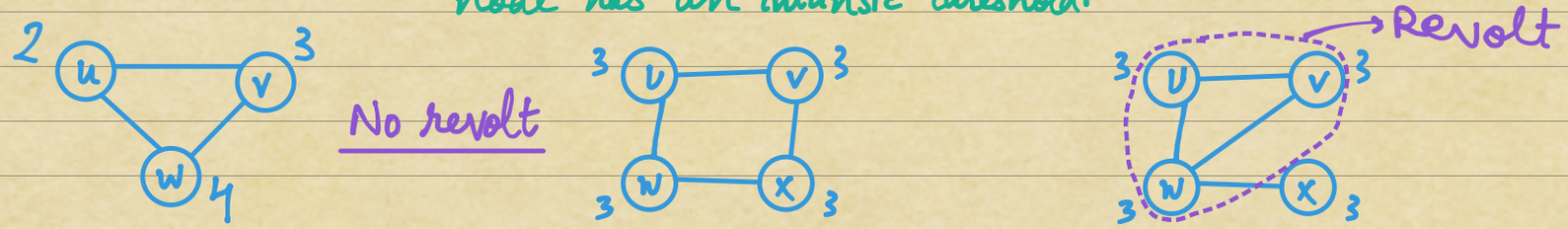
This would help in spreading the product.

→ incr the payoff to surpass the initial hesitation.

Collective Action

⊛ employees divided into communities, with very little Knowledge of each other. Advantageous for boss, as it is hard for employees to revolt.

Intrinsic Threshold :- It denotes the number of nodes needed to support a collective decision, so that a node will also support it. Each node has an intrinsic threshold.



Week 8 "Page Rank"

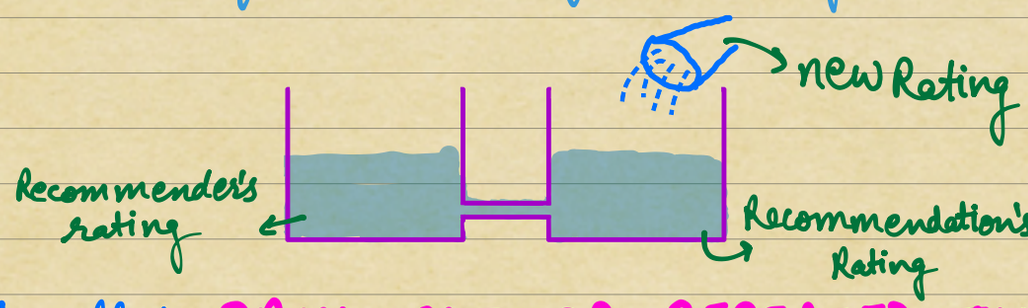
Recommender and Recommendation both have their own ratings.
Recommender's rating depends on their Recommendations and
Recommendation's rating depends on their Recommenders.

Beach and Book store example.

Recommenders ≡ Hubs ; Recommendations ≡ Authorities

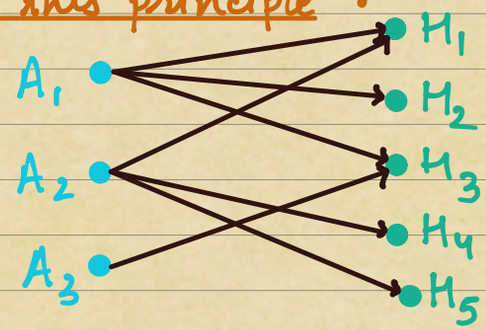
if recommender recommends something good, then both the recommender and recommendation gets better rating, here, I got a visual aid for you!

Husband wife example



It's called PRINCIPLE OF REPEATED IMPROVEMENT.

Quantifying this principle :-



$A_1 = H_1 + H_2 + H_3$

$A_2 = H_1 + H_4 + H_5$

$A_3 = H_3$

$H_1 = A_1 + A_2$
$H_2 = A_1$
$H_3 = A_1 + A_3$
$H_4 = A_2$
$H_5 = A_2$

Conclusions :-

1.) Better rated Authorities points towards better rated hubs and visa-versa.
2.) If we start with equal or unequal distribution of rating, and then if we keep redistributing the ratings, then the " of rating converge after sometime.
↳ proportion converges

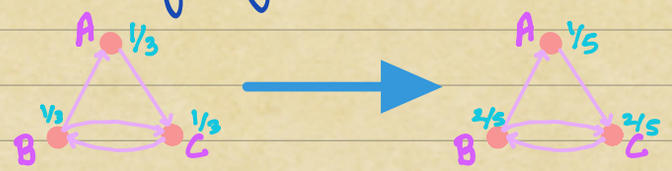
Consider the above graph to be undirected and apply equal-coin distribution method of PageRank (except that the coins are not equally distributed, and its ratings instead of coins).

Can be done in excel too!
1. distribute ratings based on pointers
2. renormalisation = $\frac{\text{rating}(A_3)}{\text{rating}(A_1+A_2+A_3)}$
3. Repeat until ratings converge

"Repeated improvement converges"

In a student recommendation network, we can tell if a student is good if: good students points to them and if they points good students (Recurssive def.)

Revisiting PageRank



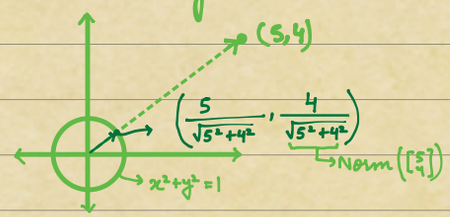
Algo:-
1.) distribute resources equally among A, B and C.
2.) redistribute the resources based on the pointer, until convergence.
↳ $A = B/2$, $B = C$, $C = A + B/2$

law of conservation of resources ⇒ $A+B+C = \frac{B}{2} + C + A + \frac{B}{2} = A+B+C$

Understanding Convergence

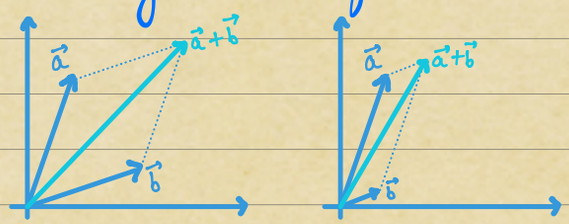
Pre-requisites

Normalising a vector



We use it when we only care about the direction and not the magnitude of the vector.

Parallelogram law of vector addition



if $|\vec{a}| >> |\vec{b}|$, then direction of $\vec{a}+\vec{b}$ will be almost equal to the direction of $\vec{a}$

So, we can conclude:-

$$A\vec{v} = A(\lambda_1\vec{v}_1 + \lambda_2\vec{v}_2) = \lambda_1 A\vec{v}_1 + \lambda_2 A\vec{v}_2$$

$$A\vec{v} = \lambda_1^2\vec{v}_1 + \lambda_2^2\vec{v}_2$$

$$\therefore A^k\vec{v} = \lambda_1^k\vec{v}_1 + \lambda_2^k\vec{v}_2 \quad (\text{Let, } \lambda_1 > \lambda_2)$$

eigenvalues are usually distinct

Repeated matrix multiplication and normalisation.

$$\underbrace{\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}}_{\text{numpy.mat('1 2; 3 4')}}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}3\\7\end{bmatrix} \xrightarrow{\text{normal}} \begin{pmatrix}0.39\\0.91\end{pmatrix} \Rightarrow \begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\begin{bmatrix}0.39\\0.91\end{bmatrix} = \begin{bmatrix}2.21\\4.81\end{bmatrix} \ldots \text{so on}$$

v/numpy.linalg.norm(v)

This also converges! Maybe we can explain Pagerank convergence with this matrix convergence!

This convergence happens for any random matrix and any random vector.

Eigenvectors and Eigenvalues

$A\vec{v} = \lambda\vec{v}$ ($\vec{v}$ doesnot changes its span after the transformation)

eigenvector, eigenvalue

Usually, for a 2x2 matrix there exsists two <u>linearly independent</u> eigenvectors

↳ 2 L.I. vectors span the entire plane

i.e. $z = \alpha v_1 + \beta v_2$; $z \in \mathbb{R}^2$ and v_1 and v_2 are L.I.

interesting fact

if $x < y$ and $z = \left(\frac{x}{y}\right)^k$, then as $k \to \infty$ $z \to 0$ or $y^k >> x^k$

Connecting the dots

$$A^k\vec{v} = \lambda_1^k\vec{v}_1 + \lambda_2^k\vec{v}_2 \quad (\vec{v}\in\mathbb{R}^2 \text{ and } \lambda_1 > \lambda_2)$$

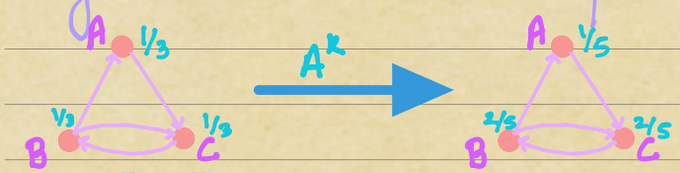
as $k\to\infty$, $\lambda_1^k >> \lambda_2^k$, $\lambda_1^k\vec{v}_1$ is a really big vector as compared to $\lambda_2^k\vec{v}_2$,

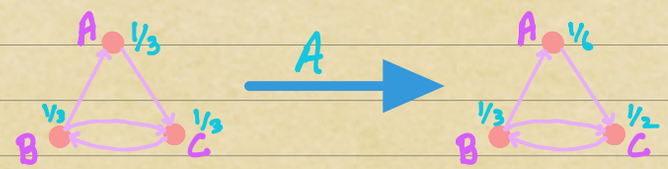
So, $\lambda_1^k\vec{v}_1 + \lambda_2^k\vec{v}_2 \approx \lambda_1^k\vec{v}_1$

$$\therefore A^k\vec{v} \approx \lambda_1^k\vec{v}_1 \quad \text{for all } \vec{v}\in\mathbb{R}^2$$

"By repeatedly apply A to any random vector $\vec{v}$, we converges to a vector in the direction of eigenvector of A."

PageRank ⟿ Matrix operation





$$\vec{v} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

direction of eigenvector of A = direction $\left(\begin{bmatrix} 1/5 \\ 2/5 \\ 2/5 \end{bmatrix}\right)$

$$\begin{array}{c|ccc} & A & B & C \\ \hline A & 0 & 1/2 & 0 \\ B & 0 & 0 & 1 \\ C & 1 & 1/2 & 0 \end{array} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} = \begin{bmatrix} 1/6 \\ 1/3 \\ 1/2 \end{bmatrix}$$

A

proportion of A's resources that goes to C.

$$A\vec{v} = \lambda\vec{v}$$

$$(A - I\lambda)\vec{v} = 0 \Rightarrow \det(A - I\lambda) = 0$$

$$\det\left(\begin{bmatrix} -\lambda & 1/2 & 0 \\ 0 & -\lambda & 1 \\ 1 & 1/2 & -\lambda \end{bmatrix}\right) = -\lambda\left(\lambda^2 - \tfrac{1}{2}\right) - \tfrac{1}{2}(-1) = 2\lambda^3 - \lambda - 1 = (\lambda-1)(2\lambda^2 + 2\lambda + 1) = 0$$

$$\lambda = 1,\ -\frac{1+i}{2},\ -\frac{1-i}{2}$$

$$\begin{bmatrix} 0 & 1/2 & 0 \\ 0 & 0 & 1 \\ 1 & 1/2 & 0 \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 1 \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \Rightarrow \begin{cases} v_2/2 = v_1 \\ v_3 = v_2 \\ v_1 + v_2/2 = v_3 \end{cases} \Longrightarrow$$

eigenvectors $= (x/2, x, x) \; \forall \; x \in \mathbb{R}$

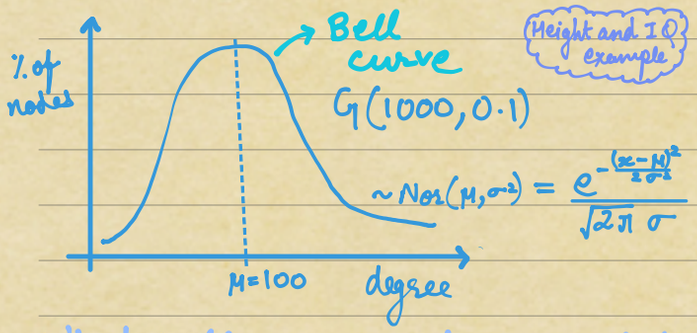
normalised eigenvector $= (0.2, 0.4, 0.4)$

<u>Markovian matrix</u> : Every column sums to 1.

Highest eigenvalue of such a matrix is 1. $(\therefore A^k \vec{v} = 1 \cdot \vec{v})$

Week 9 "Rich gets Richer phenomenon"

Normal distribution

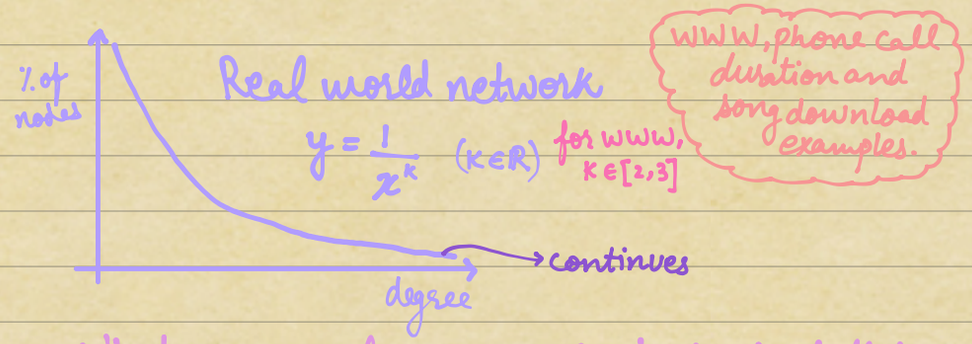


What bell curve is trying to tell is that, there are very few nodes with very high and very low degree, and most have an average degree.

It appears when the quantity is a sum of random variables. (CLT)

eg:- take 10 random nos. in [1, 100] and add them; it is very unlikely to get a 1 or a 1000 and very likely to get a 505.

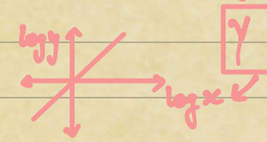
Power law distribution



What power law curve is trying to tell is that, there are very few nodes with very high degree and most of the nodes with very low degree.

If log log plot is a str. line ⇒ power law distrib.

$y = \frac{1}{x^k} \Rightarrow \log y = \log 1 - k \log x$

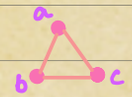


$\boxed{Y = -k\,X}$ → Str. line on a log log plot

It appears because of preferential attachment Hypothesis. (most convincing hypothesis, and we don't have any proof of it)

Prefexential attachment (example)

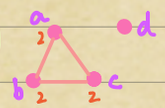
Initially, there are 3 friends in the class.



Rules of new friendships:-

1. Every new coming student makes strictly 2 friends.
2. The new student is attracted more to someone with many friends and less to someone with less friends. (i.e. based on the degree distribution)

1.

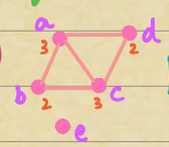


$p(d-a) = \frac{2}{6}$ → degree of 'a'; 6 → sum of degrees = 2×(no. of edges)

$p(d-b) = \frac{2}{6}$

$p(d-c) = \frac{2}{6}$

2.



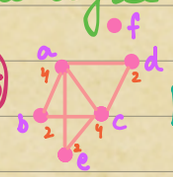
$p(e-a) = \frac{3}{10}$

$p(e-b) = \frac{2}{10}$

$p(e-c) = \frac{3}{10}$

$p(e-d) = \frac{2}{10}$

3.



$p(f-a) = \frac{4}{14}$

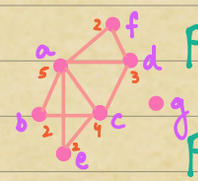
$p(f-b) = \frac{2}{14}$

$p(f-c) = \frac{4}{14}$

$p(f-d) = \frac{2}{14}$

$p(f-e) = \frac{2}{14}$

4.



$p(g-a) = \frac{5}{18}$

$p(g-b) = \frac{2}{18}$

$p(g-c) = \frac{4}{18}$

$p(g-d) = \frac{3}{18}$

$p(g-e) = \frac{2}{18}$

$p(g-f) = \frac{2}{18}$

Observation ⇒ If we do this for 1000 students, then we will see, the students with lots of friends will keep getting more and more friends. So, there will be very few students with lots of friends, and a lot of students with very few friends. POWER LAW!!

preferential attachment aka Rich gets Richer phenomenon aka Mattew effect aka Barabasi-Albert Model

Aditi Juneja

Week 10 "Disease model"

Rich get richer, but how does people become rich in the first place? How much of the popularity is because of the talent and how much is because of RGR phenomenon? How to remove the noise by RGR pheno. and see the true popularity?

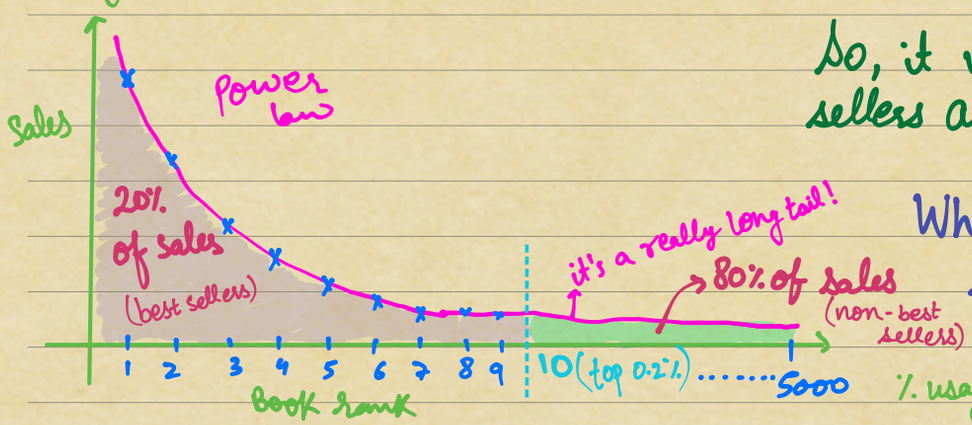
Experiment Realease same bunch of songs on different portals for download. and everyone gets to see the no. of downloads for every song, and a person can see and download from, one portal only.

Observation The download distribution, for all portals, follows power law distribution. But, the Songs' rank were not same for all portals. (no. of downloads must have influenced the new people) But, the top songs were same for all portals.

Rich gets richer phenomenon also happens on Quora, StackExchange, youtube etc.

Importance of non-popular entities (Book store example)

Why do book stores have so many books? Why don't they just have best sellers?



So, it would be stupid to remove all non-best sellers as 80% of sales comes from them. (This also happen in youtube)

When long tail, niche things dominates the sales, it's called long tail phenomenon.

(Zipf's law)

% usage of a word and word rank (defined based on the frequency of a word) also gives a power law distribution

A Mind blowing fact :- Over time, best sellers become non-best sellers, and new best sellers take their places, but the above curve remains the SAME!! 🤯

Epidemics

Difference between spread of ideas and spread of a disease :-

1.) We can <u>choose</u> to adopt an idea/habit/product, but we cannot choose to get or not get a disease (although we can determine the probability based on the immunity of the population wrt the disease)

2.) It's easier to track a social contagion on a network, than to track an actual contagion.

Spread of disease depends on:-

1.) Degree of contagiousness (of the pathogen)

2.) Network density (density $\propto$ r/o spread and density depends on the type of disease i.e. air-born, vector-born, STDs etc.)

If we have a tree shaped network, with each node having K children, and 'p' being the probability of infection, then the E[infected people at i^{th} level] = $p^i K^i$

probability of someone at i^{th} level, getting infected.

people at i^{th} level

$$E[\text{infected people at } i^{th} \text{ level}] = (pK)^i$$

Give (p, K), how can we predict if epidemic will happen or not?

$$f(p, K) = \begin{cases} 1, & \text{epidemic} \\ 0, & \text{no epidemic} \end{cases}$$

What is f?

pK = Basic Reproductive no. (R_0) = E[Secondary infections] R_0 can be used as a measure of immunity wrt disease.

↳ i.e. #infections passed on by a single node

$R_0 < 1 \Rightarrow$ secondary infections < 1 (disease will most probably in first level)

No epidimic ; P(Epidemic) = 0

$R_0 > 1 \Rightarrow$ disease will persist with a +ve probability of an epidemic. P(Epidemic) > 0

(Very unlike that the disease will die.)

knife edge property

To stop an epidemic we need to bring R_0 below 1, and that can be done by:

1.) Reducing p, i.e. by taking precautions, being hygienic etc. , or/and

2.) Reducing K, i.e. by isolation, masks etc.

$T_i \Rightarrow$ time period of illness ; $p \Rightarrow$ probability of getting infected by an ill neighbor

Non-tree shaped networks disease spreading models :-

1.) **SIR**

Susceptible $\longrightarrow$ Infected (T_i) $\longrightarrow$ Recovered (permanent immunity)/Removed (dead) (eg :- Measles)

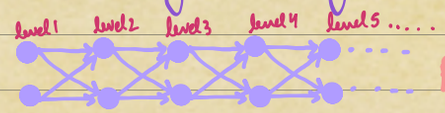
The process will end, eventually. (all nodes will become either recovered/removed or a combination of recovered and susceptible)

2.) **SIS**

Susceptible $\rightleftarrows$ Infected (T_i) (eg :- common cold)

The process may or may not end (all nodes will be susceptible or it never ends)

Infinite Network



$p = 2/3$ $R_0 = \underbrace{2}_{\text{2 children of each node}} \times \frac{2}{3} = \frac{4}{3} > 1$

$P(\text{going to level 2}) = 1 - \left(\frac{1}{3}\right)^4 \approx 0.98$, $P(\text{going to level } x) = 1 - \frac{1}{3^{2x(x-1)}}$ $x \to \infty$, $P \to 1$ $\therefore$ disease persists forever?

Percolation Model (static version of disease model, without T_i ; or if you only want to know about a node)

Think of the network as a pipeline system, with edges as pipes and with probability 'p' an edge is closed pipe.

If there is path of open pipes from the initial infected node to some node, then that node will get infected during the process.

Analysing R_0

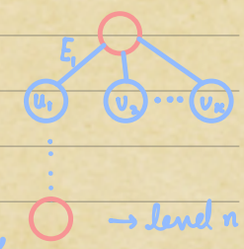
$q_n = P(\text{at least 1 person is infected at level } n \text{, of a tree network})$; $q_0 = 1$; $q^* = \lim_{n\to\infty} q_n$

To prove :- if $R_0 < 1$ then $q^* = 0$, and
if $R_0 > 1$ then $q^* > 0$

Proof :-

$*_1 \Rightarrow$ the infection reaches the nth level from the root, infected node and u_1

$$P(*_1) = \underbrace{P(\text{infection passed through edge } E_1 \text{ to node } u_1)}_{p} \times \underbrace{P(\text{infection goes from } v_1 \text{ to } n^{th} \text{ level})}_{q_{n-1}}$$



$$P(*_1) = pq_{n-1}$$

$$q_n = 1 - P(\text{all } * \text{ fail}) = 1 - (1 - pq_{n-1})^k$$

$$q_1 = 1-(1-pq_0)^K = 1-(1-p)^K$$
$$q_2 = 1-(1-p(1-(1-p)^K))^K$$

Let, $f(x) = 1-(1-px)^K$

$\Rightarrow q_1 = f(q_0)$

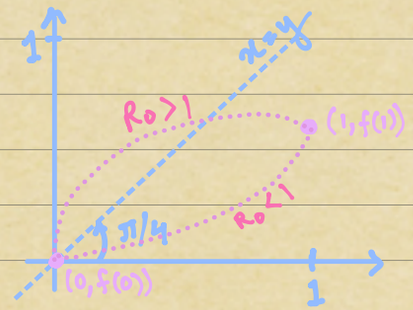
$\Rightarrow q_2 = f(f(q_0))$

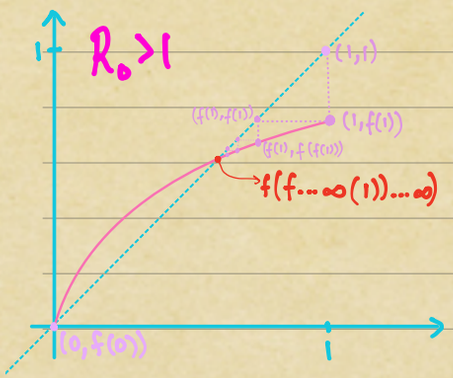
$$\therefore \; q_* = f(f \cdots \infty (q_0))) \cdots \infty)$$
$$q_* = f(f \cdots \infty (1))) \cdots \infty)$$

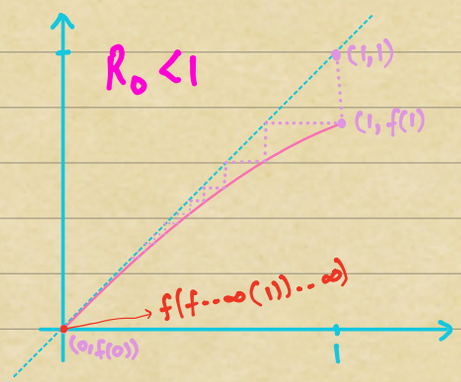
Plotting $f(x)$

$f(0)=0$

$f(1) = 1-(1-p)^K < 1$







q_* = intersection of $f(x)$ and $x=y$

$q_* = 0$

finding slope

$$f'(x) = \frac{d(1-(1-px)^K)}{dx} = pK(1-px)^{K-1}$$

$f'(0) = pK = R_0$ (slope at origin)

if $R_0 > 1$ slope $= \tan\theta > 1 \Rightarrow \theta > \frac{\pi}{4}$

if $R_0 < 1$ slope $= \tan\theta < 1 \Rightarrow \theta < \frac{\pi}{4}$

R|0 change of R|0 change

$$f''(x) = \frac{d(pK(1-px)^{K-1})}{dx}$$
$$= -p^2K(K-1)(1-px)^{K-2}$$
$$f''(0) = -p^2K(K-1) < 0$$
$$(K>1)$$

Aditi Juneja

Week 11 "Small World Phenomenon"

Is the world connected? What is the average shortest path b/w a pair of people?
↳ yes

Any 2 people on the world friendship network are 6 hops (or 5 people) away from each other, on an average!!

Experiment

Milgram's observation (1967)

pass on a letter from on person to a far away person and everyone passes the letter to a friend that they think might be closest to the target node.

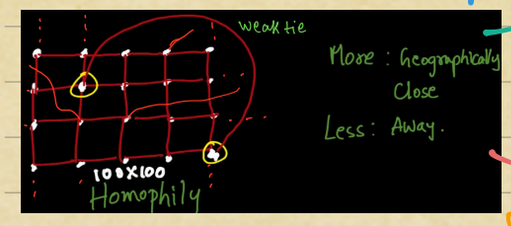
It's counter intuitive bcoz in a world of 7 B people, the diameter is just 6.

It's intutive ↳

100 people I know
10000 people: 2nd level
100^8 people → 8th level

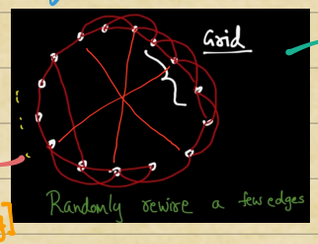
Watts-Strogatz Model (1998) (email instead of letters)

Conclusion:- S.W.P. happens because of Homophily and weak ties. (Hypothesis)



→ world as a 2D grid

→ Generative models [Homophily + Rewiring]



→ World as a circle

Rewire:- remove and randomly put the edge somewhere else, with some probability = p

We can use Milgram's Model as a Search Algo.!!

(local/greedy algo.)

1.) Make a generative model

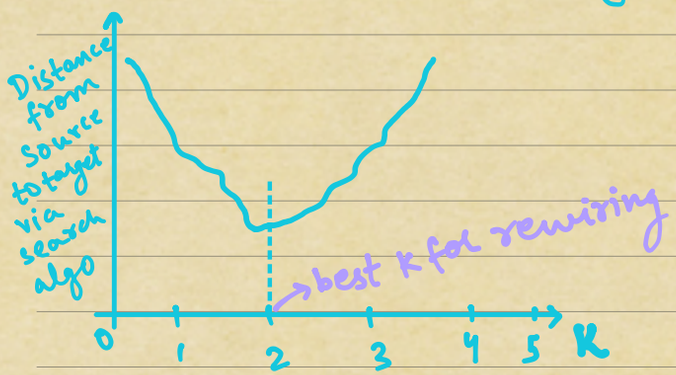
→ geographically close people are connected

→ for weak ties, $p(u-v) = \dfrac{1}{\text{distance}(u,v)^K}$, $K = 0, 1, 2 \ldots$

→ clustering exponent :- measure of how close a network is to a clique (↳ complete graph)

metro/bus route modelling examples

• Source • target (in a grid)



⊛ for 1D model (ring), best value of K = 1

2.) <u>Decentralised Search</u> :- go to the node i.e. closest to target node, geographically unless we reach a neighbor of target. (aka myopic search)

<u>Fun fact</u> :- The world might be a little disconnected, because there are some tribes in Northern Andaman islands, and many other areas of the world, which are very isolated from rest of the world, and they don't want others to come near their place.

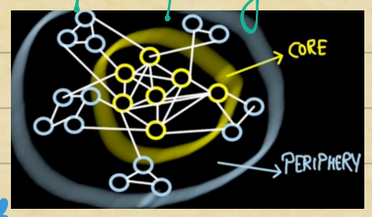
Aditi Juneja

Week 12 "Targeting Key nodes"

Virality of an entity, on a network depends on:-

1.) Quality of the entity
2.) Structure of the network
3.) Key nodes' opinions

Core-periphery structure



To make an entity viral, infect the core node with it.

Defining 'cores' :- n^{th} core is a subgraph, in which every node have a degree $\geq n$.

nx.core_number(G)

Finding 'cores' :- K-shell decomposition Algo.
↳ aka bucket

1.) counter = 1 , buckets = [ ], graph = nx.read_gml('web.txt')
2.) Repeat steps 3 to 7, until graph != empty
3.) bucket = [ ], deg_dist = get_degrees(graph)
4.) while there exsists a node with degree ≤ counter in graph.nodes(), repeat step 5
5.) for i in graph.nodes(), if degree(i) ≤ counter, ① Add i to bucket, and
② Remove i from graph
6.) Add bucket to buckets
7.) counter += 1

K-core = Buckets[K] ∪ Buckets[K+1] ∪ . . . Buckets[n]

$$K\text{-core} = \bigcup_{j \geq K} \text{Buckets}[j]$$

Influential power If {1, 2, 3} are infected by an entity in the network, then they independently infects their neighbors with 'p' probability, and at the end of one iteration, infected set = {1, 2, 3, 7, 9, 4, 6}, then the influencial power of {1, 2, 3} = len (infected set) = 7.

* If we compare 4 centrality measures (i.e. degree, betweenness, closeness and coreness) with the influence power, for each node, then we observed that coreness and influential power are most related

Pseudo core

"Cascade Capacity of a shell" = I.P. (set(nodes in the shell)).

* Sometimes hard to infect the core, due to financial reasons etc. in that cases we can infect a no. of pseudo cores, which results in maximum cascade.

Aditi Juneja