

# Social Media Analytics (SMA)

## *Community Detection*

### *Part 1*

**Marco Viviani**

University of Milano-Bicocca

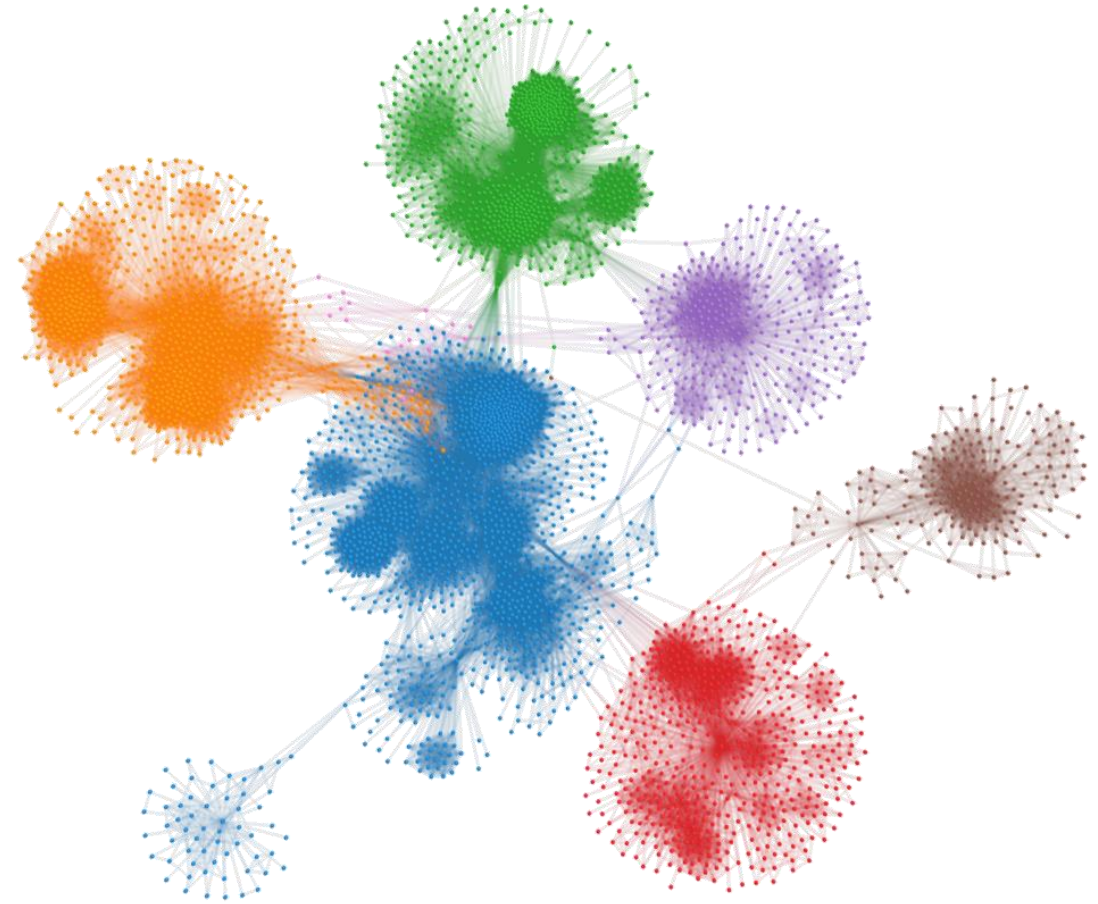
*Department of Informatics, Systems, and Communication*



DIPARTIMENTO DI  
INFORMATICA, SISTEMISTICA E  
COMUNICAZIONE

# Outline

- Some terminology
  - Cohesive subgroup
  - Community
- Community characteristics
- Aim of analyzing communities
- Some related concepts
  - Scale-free network
  - Preferential attachment
  - Assortativity



# Cohesive Subgroup

- **Cohesive subgroups** typically refer to small, tightly connected groups of nodes within a network.
- These subgroups are characterized by **strong internal connections** and are often identified through various network analysis methods, such as graph theory or clustering algorithms.
- Cohesive subgroups **may not** necessarily imply that they **represent significant or meaningful divisions** within the network.
  - They can be small, isolated groups of nodes that are highly interconnected for various reasons.

# Community

- The term **community** is often used to describe larger, more significant, and **functionally meaningful** divisions within a network.
  - Communities are typically composed of nodes that **share common characteristics, interests, functions, or roles** within the network.
- The concept of community is often used to analyze and understand the **organization, function, and dynamics of complex systems**, including social networks, biological networks, and many other real-world systems.
- The term **cohesive subgroup** is more general and may not always imply a strong functional or thematic relationship among the nodes in the subgroup.

# Key Characteristics of Communities (1/2)

- **High Internal Connectivity:** Nodes within a community are strongly connected to each other.
  - Meaning there are many edges (links) between nodes within the same community.
- **Low External Connectivity:** Nodes in a community have fewer connections to nodes outside of their community.
  - The connections to nodes outside the community are typically weaker or less frequent.

# Key Characteristics of Communities (2/2)

- **Modular Structure:** Complex networks often exhibit a modular or hierarchical structure.
  - Multiple levels of communities within communities.
  - This nested structure can reveal insights into the organization of the network.
- **Functional Significance:** Communities may represent groups of nodes that have a similar function, role, or purpose within the network.
  - For example, in a social network, a community could correspond to a group of friends with common interests.

# Observable VS Latent Communities

- In the context of social media and online social networks, there are two types of communities that are often discussed: **observable communities** and **latent communities**.
- **Observable communities**: those that are explicitly visible or easily identifiable based on the available data and interactions within a social network.
  - These communities can be directly observed or measured using data such as user connections, interactions, interests, and affiliations.
- **Latent communities**: not explicitly defined by user actions or interactions but are inferred through data analysis and machine learning techniques.
  - Latent communities are typically discovered using algorithms that aim to find patterns, similarities, or hidden groupings among users.

















# Observable Communities

- **Explicit user groups**: some social media platforms allow users to create or join groups or pages focused on specific topics, interests, or affiliations.
- **Hashtags and topics**: communities can also be defined by the use of common hashtags or engagement with specific topics.
- **Friendships and connections**: the connections between users, such as friendships on Facebook or followers on Twitter, can be used to identify observable communities.
  - For example, a friend network on a platform like Facebook can reveal communities of real-world friends and acquaintances.



# Observable Communities: Examples

	Name: <b>Social Computing</b> Type: Organizations Members: 14 members
	Name: <b>Social Computing</b> Type: Internet & Technology Members: 12 members
	Name: <b>Social Computing Magazine</b> Type: Internet & Technology Members: 34 members
	Name: <b>Trustworthy Social Computing</b> Type: Internet & Technology Members: 28 members
	Name: <b>Social Computing for Business</b> Type: Internet & Technology Members: 421 members
	Name: <b>UCLA Social Sciences Computing</b> Type: Internet & Technology Members: 22 members

	<b>I * Urban LIFE in Metropolis ///</b> 4,286 members   31 discussions   89,645 items   Created 46 months ago   <a href="#">Join?</a> UrbanLIFE, People, Parties, Dance, Musik, Life, Love, Culture, Food and Everything what we could imagine by hearing that word URBANLIFE! Have some FUN! Please add... ( <a href="#">more</a> )	
	<b>Islam Is The Way Of Life (Muslim World)</b> 619 members   13 discussions   2,685 items   Created 23 months ago   <a href="#">Join?</a> The word islām is derived from the Arabic verb aslama, which means to accept, surrender or submit. Thus, Islam means submission to and acceptance of God, and believers must... ( <a href="#">more</a> )	
	<b>* THE CELEBRATION OF ~LIFE~ (Post1~Award1) [only living things]</b> 4,871 members   22 discussions   40,519 items   Created 21 months ago   <a href="#">Join?</a> WELCOME to THE CELEBRATION OF ~LIFE~ (Post1~Award1) PLEASE INVITE & COMMENT USING only THE CODES FOUND BELOW! ☆ ☆ This group is for sharing BEAUTIFUL, TOP QUALITY images... ( <a href="#">more</a> )	
	<b>"Enjoy Life!"</b> 2,027 members   10 discussions   39,916 items   Created 23 months ago   <a href="#">Join?</a> There are lovely moments and adorable scenes in our lives. Some are in front of you, and some are just waiting to be discovered. A gaze from someone we love, might touch the... ( <a href="#">more</a> )	
	<b>Baby's life</b> 2,047 members   185 discussions   30,302 items   Created 32 months ago   <a href="#">Join?</a> This group is designed to highlight milestones and important events in your baby's life (ie 1st time smiling/crawling/sitting in a high chair/reading/playing etc). It can also be... ( <a href="#">more</a> )	
	<b>Pond Life</b> 903 members   20 discussions   6,877 items   Created 32 months ago   <a href="#">Join?</a> Pic of the week: chosen from the pool by the group admins. Nuphar by guus timpers Pond Life is a group for all aquatic flora and fauna. Koi ponds, wildlife ponds, garden ponds,... ( <a href="#">more</a> )	
	<b>Second Life</b> 10,288 members   773 discussions   257,870 items   Created 61 months ago   <a href="#">Join?</a> Welcome to the Second Life pool, the biggest group on Flickr for residents/players of Second Life, the 3D virtual world from Linden Lab. This group is not endorsed or run in any... ( <a href="#">more</a> )	
	<b>Life in Kuwait - Post 1 Award 1</b> 637 members   28 discussions   3,233 items   Created 18 months ago   <a href="#">Join?</a> About kuwait: LOCATION: Kuwait lies on the northern tip of the Arabian Gulf. It is bordered by the Kingdom of Saudi Arabia to the south and south west, and the Republic of Iraq to... ( <a href="#">more</a> )	

# Latent Communities

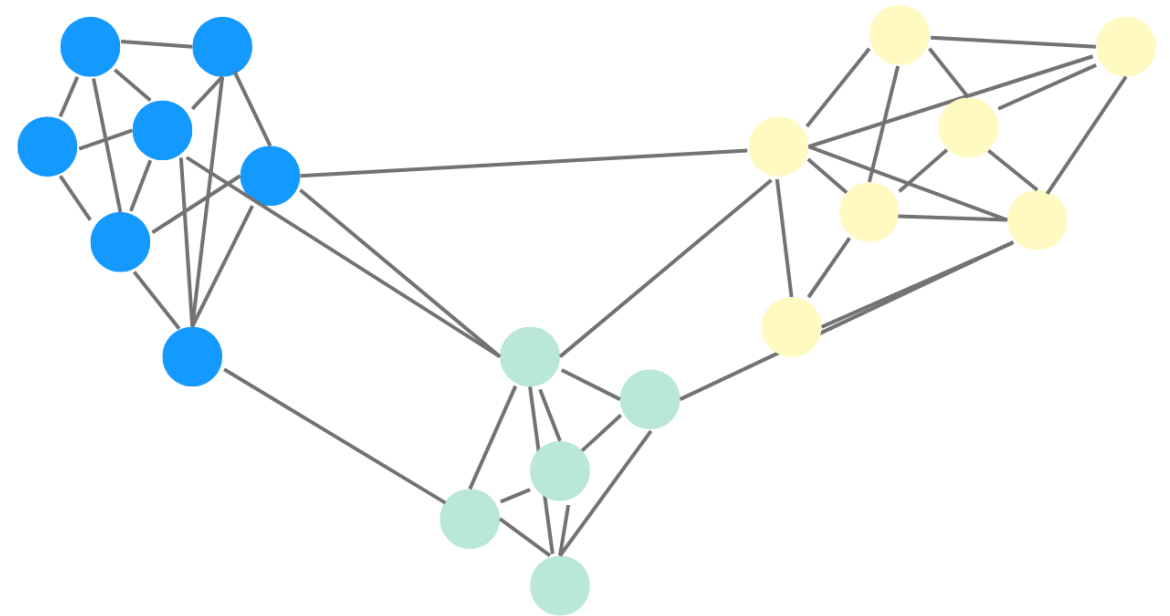
- **Network analysis:** by analyzing the structure of the social network, such as the connections between users and the patterns of interactions, researchers can identify latent communities that might not be obvious at first glance.
- **Clustering algorithms:** machine learning and network analysis techniques, like clustering algorithms, can group users based on similarities in their behavior, interests, or connections.
  - Users in the same cluster **can be** considered part of a latent community.
- **Topic modeling:** latent communities can also be inferred through topic modeling techniques that identify common themes, interests, or discussions among users.

# Why Analyzing Communities?

- **Discover functionally-related objects**
- **Study interactions between groups**
- **Infer missing node values**
- **Predict unobserved connections**

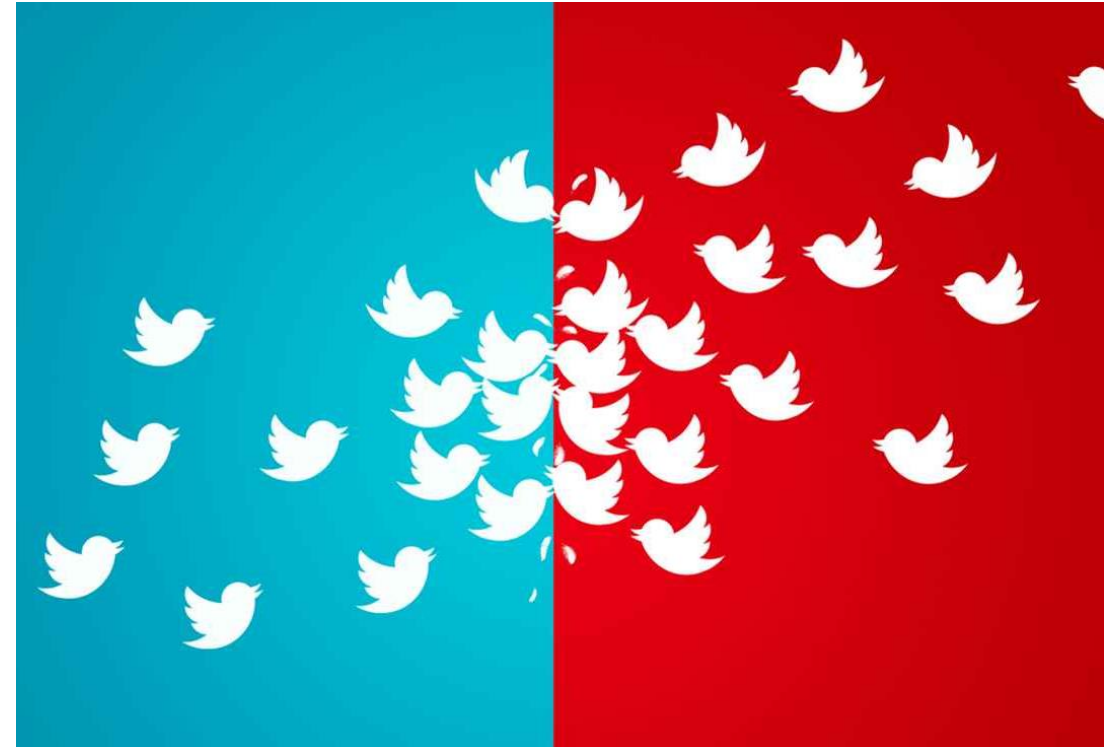
# E.g., Discover Functionally-related Objects

- Collect data on Twitter (X) interactions, such as retweets, mentions, and common hashtags, focusing on technology-related topics.
- Detect communities and analyze them to understand their function and common interests.
  - For instance, one community primarily discusses information about smartphones, while another is more focused on software development.
- To confirm that the communities represent functionally related users, the content shared by users in each community can be analyzed.
  - Social Content Analysis → Next lectures.



# E.g., Study Interactions between Groups

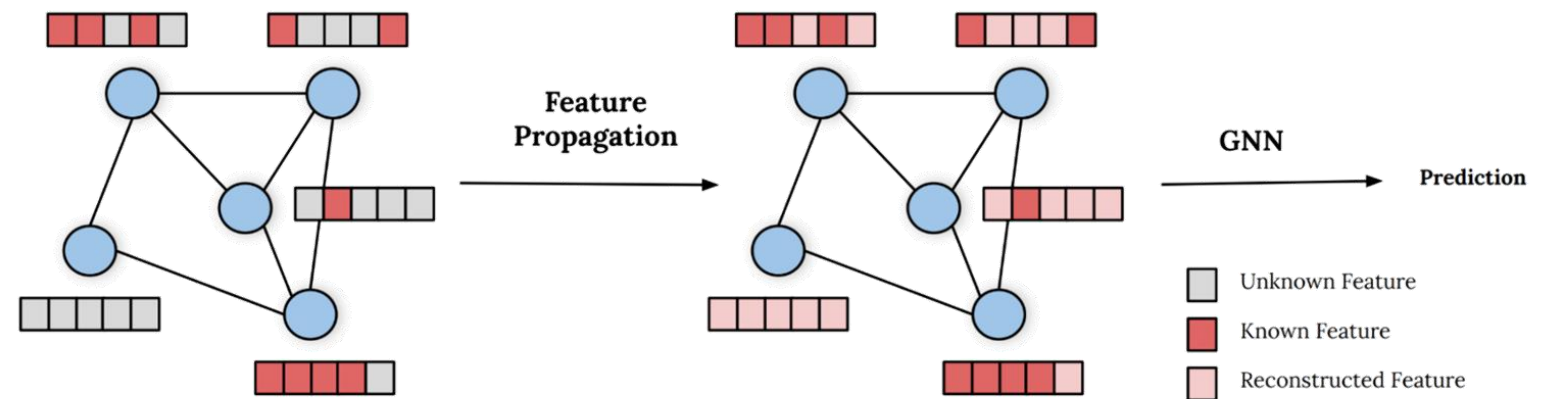
- Collect a dataset of tweets related to a politically polarized topic, i.e., a contentious election or a controversial policy issue.
- Apply community detection algorithms to identify clusters of users within the network.
  - These clusters often represent echo chambers or ideological groups.
- Perform sentiment analysis on the tweets in each community to determine the prevailing sentiment (positive or negative) associated with the political discussion.
  - Social Content Analysis → Next lectures.



# E.g., Infer Missing Node Values

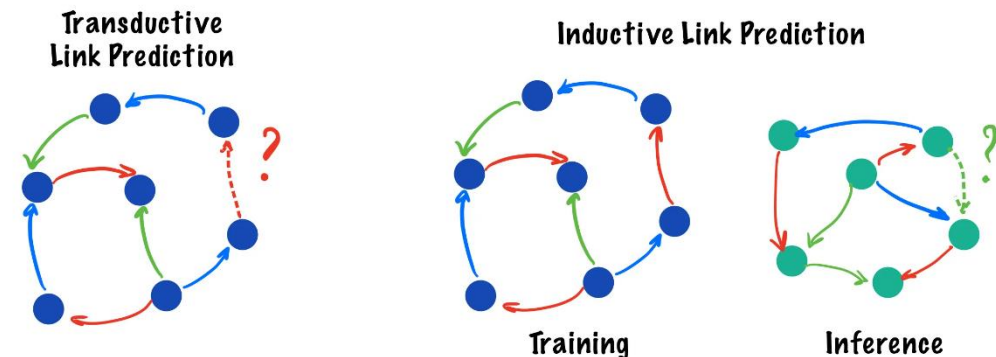
- **Gather a dataset** of users in a social network.
  - For each user, information about their connections (friends/followers), interests (e.g., liked pages), and some demographic information (e.g., age).
- **Create features** that can be used to predict users' ages (E.g., number of connections, frequency of posts, etc.)
- **Use a trained supervised model** to predict the ages of users in a test set, where age information is missing.
- **Analyze the model's predictions** to understand its accuracy and the quality of age predictions for users with missing age values.

[https://blog.twitter.com/engineering/en\\_us/topics/insights/2022/graph-machine-learning-with-missing-node-features](https://blog.twitter.com/engineering/en_us/topics/insights/2022/graph-machine-learning-with-missing-node-features)



# E.g., Predict Unobserved Connections

- **Collect data on researchers in an academic network**, including information on their research interests and previous collaborations.
  - Researchers are represented as nodes, and previous collaborations are represented as edges in the network.
- **Apply community detection** algorithms to identify research communities with shared interests or similar research areas.
  - These communities are often inferred from co-authorship networks or citation patterns.
- **Create features** for each researcher based on their individual attributes, such as research interests and community membership, and features related to their potential collaborators within and outside their community.
- **Train a model** using the training set, with known collaborations as the target variable and the engineered features as input.
  - The model learns to predict new collaborations based on shared interests and community structure.



<https://towardsdatascience.com/inductive-link-prediction-in-knowledge-graphs-23f249c31961>

# Some Related Concepts

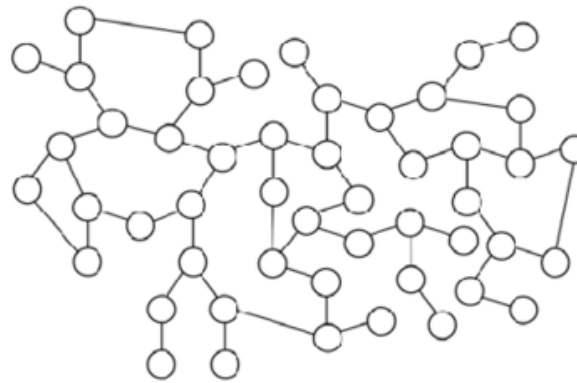
- **Scale-free network**
- **Preferential attachment**
- **Assortativity**



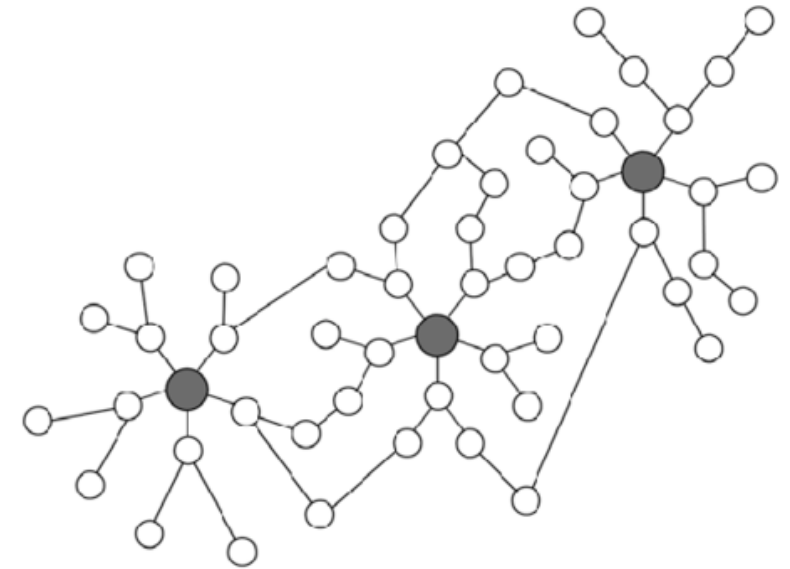
# Scale-free Network

- A **scale-free network** is a type of complex network characterized by a specific degree distribution.
- In a scale-free network, a few nodes (**hubs**) have a significantly higher degree (i.e., more connections) than the majority of nodes.
- This leads to a **power-law degree distribution**, where a small number of nodes have many connections while most nodes have only a few.

## Scale-free Network: Example



(A) Random network

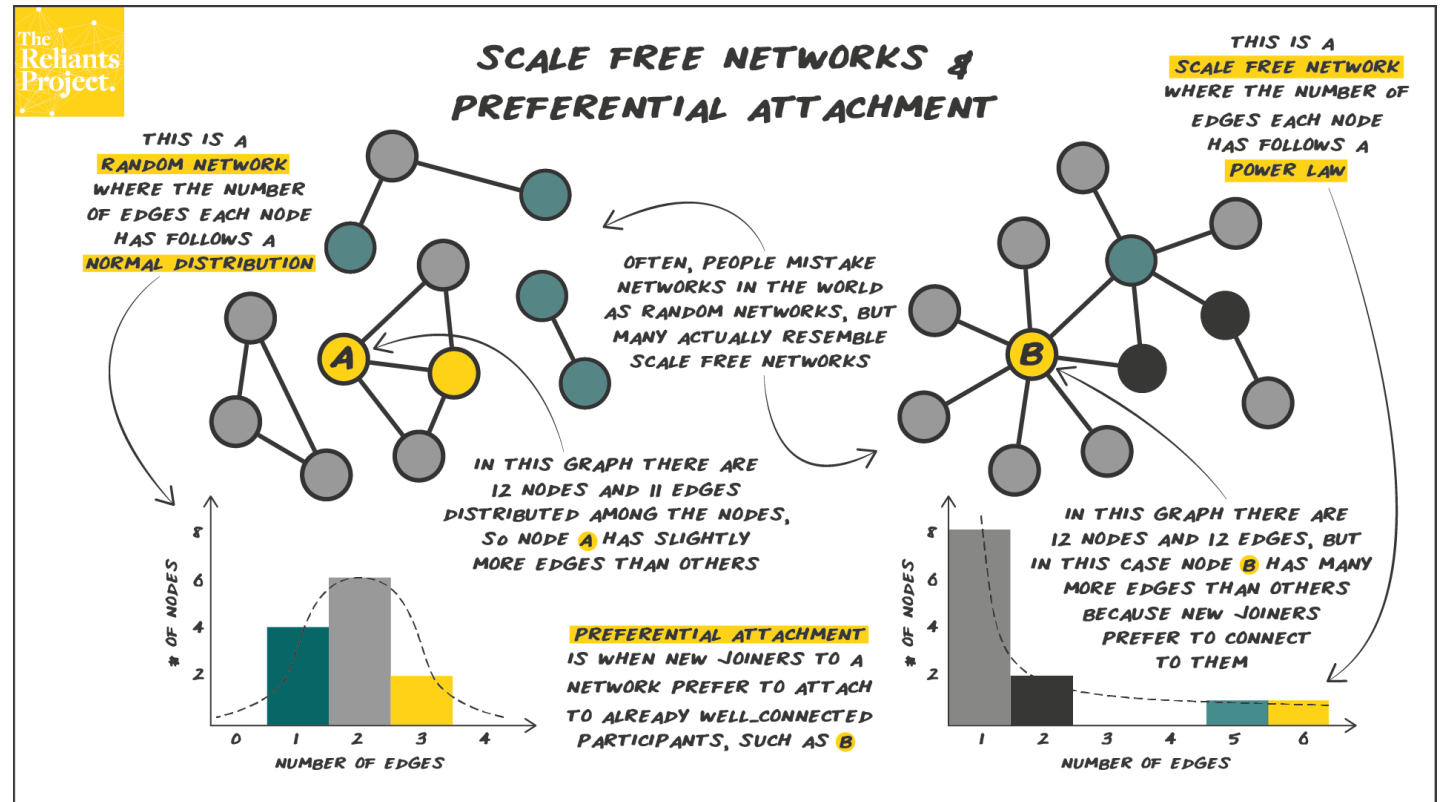


(B) Scale-free network

# Preferential Attachment

- **Preferential attachment** is a mechanism that explains the formation of scale-free networks.
- It suggests that new nodes in a network are **more likely to connect to existing nodes that already have high degrees**.
  - In other words, nodes with more connections are more attractive to new nodes joining the network.
- This process **reinforces the growth of hubs** and results in the scale-free property.

# Preferential Attachment: Example



# Preferential Attachment: Why?

- This phenomenon can be driven by several factors, including **popularity**, **quality**, and a **mix of both**.



# Popularity-Based Preferential Attachment

- In some cases, nodes in a network become more attractive and accumulate more connections simply because they are already popular or well-connected. This is sometimes referred to as the "**rich get richer**" principle.
- Some reasons for popularity-based preferential attachment include:
  - **Visibility**: popular nodes are more visible and easily discoverable, making them more likely to receive new connections.
  - **Social proof**: people tend to trust and follow nodes that already have many connections or followers, assuming that they provide valuable information or resources.
  - **Network effects**: in some networks, the value of being connected to a popular node increases as more users connect to it, reinforcing its popularity.

# Quality-Based Preferential Attachment

- In other cases, nodes gain more connections because they consistently produce **high-quality content, services, or products**.
- Reasons for quality-based preferential attachment include:
  - **Content excellence**: nodes that consistently produce high-quality content or services attract more connections because users appreciate the value they provide.
  - **Trustworthiness**: quality nodes are trusted sources of information or services, which encourages more users to connect to them.
  - **Longevity**: nodes that maintain high quality over time tend to build a loyal user base, leading to further connections.

# Mixed Preferential Attachment

- In many real-world scenarios, preferential attachment is influenced by a **combination of both popularity and quality**.
  - Nodes that are both popular and provide high-quality content or services can attract connections more rapidly and maintain their influence over time.
  - Mixed preferential attachment accounts for the interplay between popularity and quality in driving the growth of connections within a network.
- **Examples**
  - Social media influencers who produce high-quality content (quality) and have a large following (popularity) often exhibit mixed preferential attachment.
  - Websites or platforms that offer both popular features and high-quality resources tend to experience mixed preferential attachment.



# Preferential Attachment and Communities

- The dynamics of preferential attachment can contribute to the **formation of communities**.
  - Nodes that are highly connected (i.e., **hubs**) tend to attract more connections over time.
  - As these hubs accumulate more links, they can serve as **natural centers** for the formation of communities.
- The detection of communities can help **identify hubs or nodes that exhibit preferential attachment** within and between communities.
- **Understanding these connections** can lead to a more comprehensive analysis of complex network structures and dynamics.
  - Communities may represent groups of nodes with different patterns of preferential attachment, and the presence of these communities can influence how nodes form connections.

# Assortativity

- **Assortativity** (or **assortative mixing**) refers to the tendency of nodes in a network to connect with other nodes that have **similar characteristics or properties**.
  - It quantifies the **correlation** between the attributes or degrees of connected nodes in a network.
  - Assortativity is also a measure of **homophily**.
- Concept introduced in 2002 by **Newman**:
  - Newman, M. E. (2002). Assortative mixing in networks. Physical review letters, 89(20), 208701. <https://doi.org/10.1103/PhysRevLett.89.208701>

# Preferential Attachment VS Assortativity

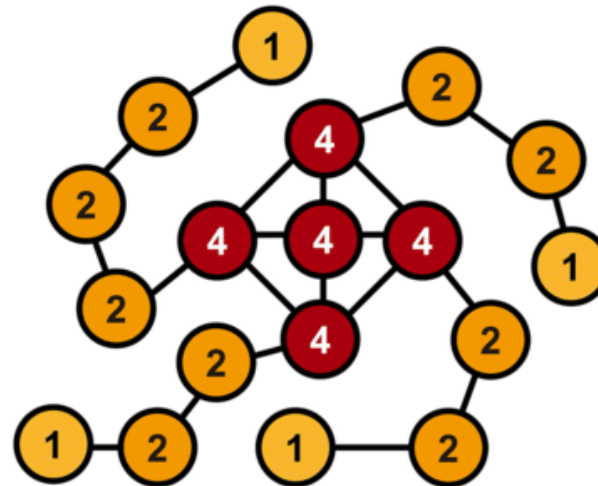
- While **preferential attachment** can result in networks with assortative mixing (positive degree correlation), it does not necessarily involve nodes having similar attributes or properties.
- Instead, **it emphasizes the structural property of nodes** gaining more connections as they become more connected.

# Assortative Networks

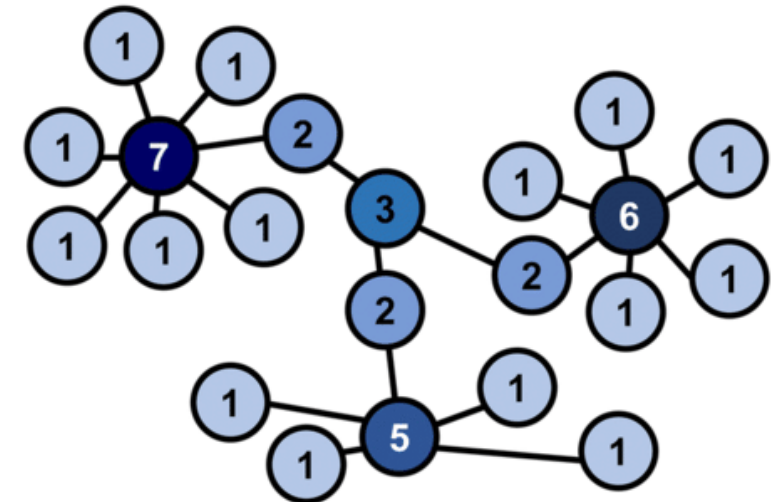
- Generally, the assortativity of a network is **determined for the degree** (number of direct neighbours) of the nodes in the network.
  - <https://doi.org/10.1093/comnet/cnv005>
- The concept of assortativity may, however, be applied to **other topological characteristics of a node** as well, such as:
  - Node weight
  - Node betweenness
  - **$k$ th level node degree** (number of nodes that can be reached in no more than  $k$  hops; also known as **expansion**)
  - ...
- In addition, assortativity may be applied to node **characteristics that are not directly topology-related**, such as related to node attributes → **homophily**.

# Assortative Networks: Example

**A** Assortative network



**B** Disassortative network



# Assortativity and Dissortativity

- Let us consider **node degree**.
- A network is said to be **assortative** when high-degree nodes are, on average, connected to other nodes with high-degree and low-degree nodes are, on average, connected to other nodes with low degree.
- A network is said to be **disassortative** when, on average, high-degree nodes are connected to nodes with low(er) degree and, on average, low-degree nodes are connected to nodes with high(er) degree.

# Assortativity, Dissortativity, and Hubs (1/3)

- By the virtue of the many links they have (based on their popularity, quality, or mixed model), **hubs are expected to link to each other.**
  - **Celebrity couples** represent a highly visible proof (in social networks) that hubs tend to date and marry each other.



# Assortativity, Dissortativity, and Hubs (2/3)

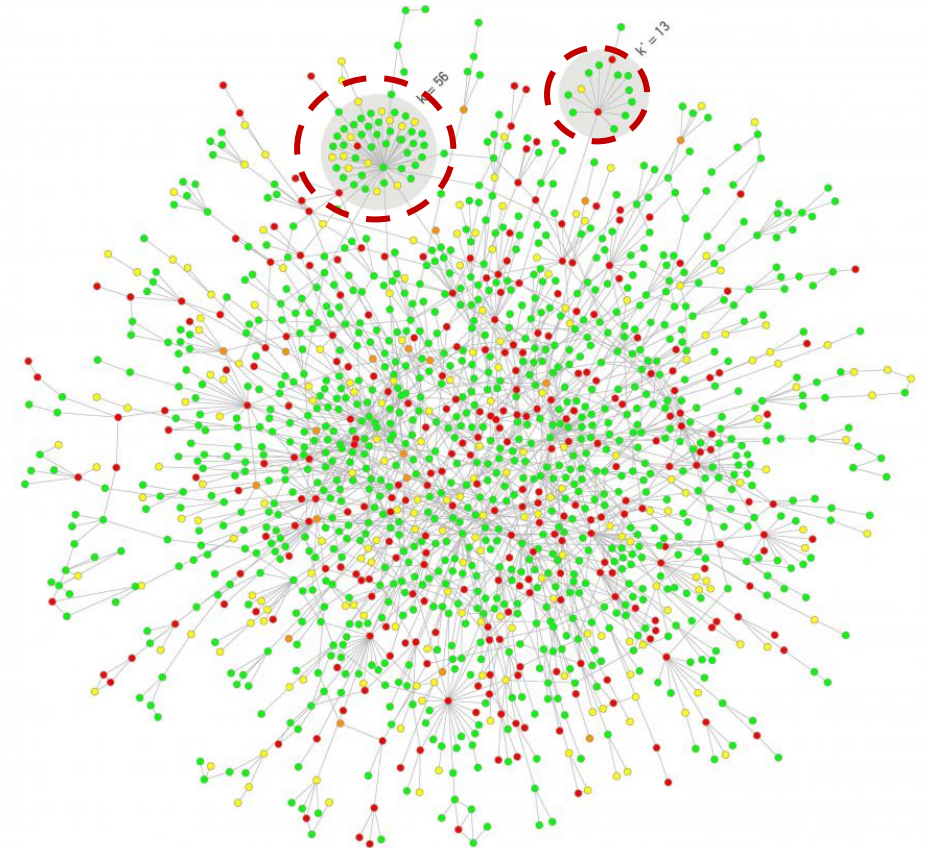
- Let's assume that a celebrity **could date anyone** from a pool of about a hundred million ( $10^8$ ) eligible individuals.
- The chances that their mate would be another celebrity from a list of 1,000 other celebrities is only  $10^{-5}$ .
  - If dating were driven by **random encounters**, celebrities would never marry each other.



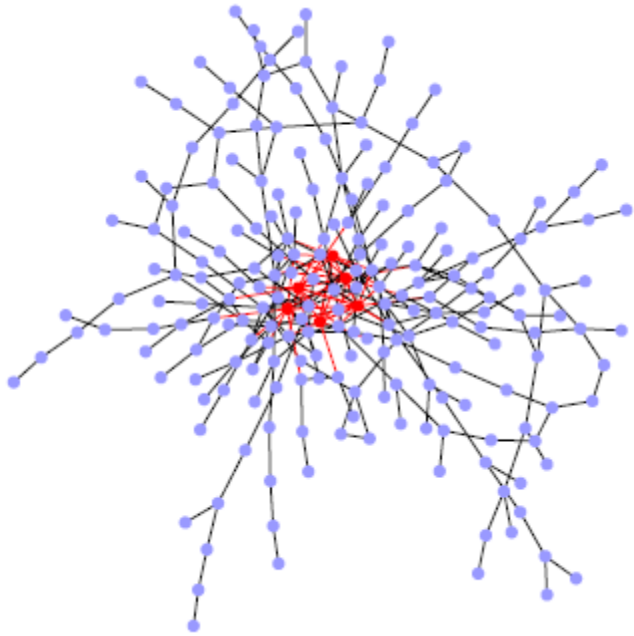


# Assortativity, Dissortativity, and Hubs (3/3)

- However, in some networks they do, **in others they do not**, and **hubs avoid hubs**!
- E.g., **Protein-protein interaction** networks.
  - This pattern is completely different from the celebrities one!
- Two **largest hubs** with degree
  - $k = 56$
  - $k' = 13$

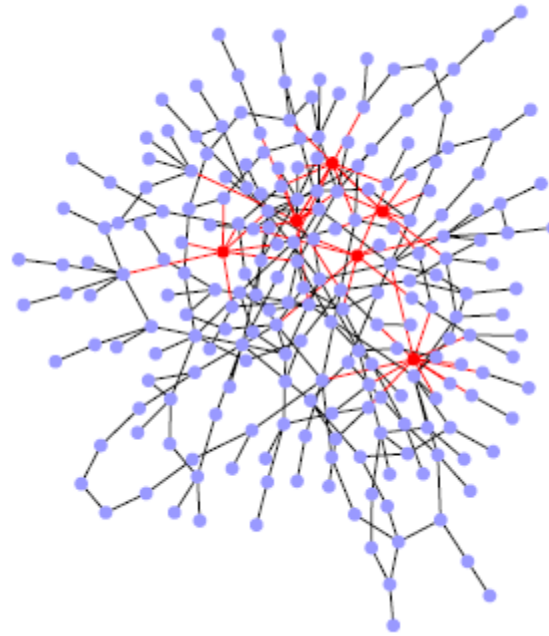


# Assortativity, Dissortativity, and Hubs (2/2)



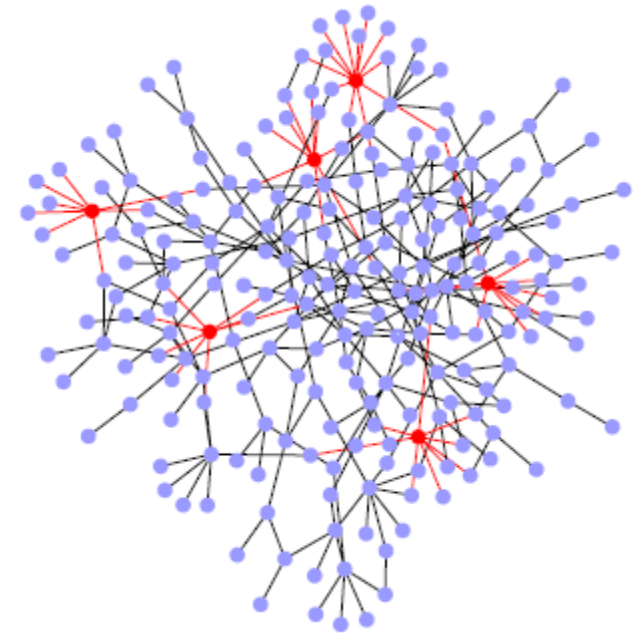
## **Assortative**

Hubs show a tendency to link to each other.



## **Neutral**

Nodes connect to each other with the expected random probabilities.



## **Disassortative**

Hubs tend to avoid linking to each other.

# Assortativity in Real Life Networks

- It is found that, **assortative networks** have been observed mostly in **social networks**, whereas many **technological and biological networks** display to be **disassortative**.
  - Newman, M. E. (2002). Assortative mixing in networks. Physical review letters, 89(20), 208701. <https://doi.org/10.1103/PhysRevLett.89.208701>
- The **generality** of assortativities in social networks has been questioned.
  - Whitney, D. E., & Alderson, D. (2010, June). Are technological and social networks really different? In Proceedings of CCS (pp. 74-81). Berlin, Heidelberg: Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-540-85081-6\\_10](https://doi.org/10.1007/978-3-540-85081-6_10)
  - Hu, H. B., & Wang, X. F. (2009). Disassortative mixing in online social networks. Europhysics Letters, 86(1), 18003. <https://doi.org/10.1209/0295-5075/86/18003>
- A variety of **recent research still states** that this is a property typical of social networks.
  - Fisher, D. N., Silk, M. J., & Franks, D. W. (2017). The perceived assortativity of social networks: methodological problems and solutions. Trends in Social Network Analysis: Information Propagation, User Behavior Modeling, Forecasting, and Vulnerability Assessment, 1-19. [https://doi.org/10.1007/978-3-319-53420-6\\_1](https://doi.org/10.1007/978-3-319-53420-6_1)

# Assortativity and Communities (1/2)

From: <https://doi.org/10.1002/cpe.4040>

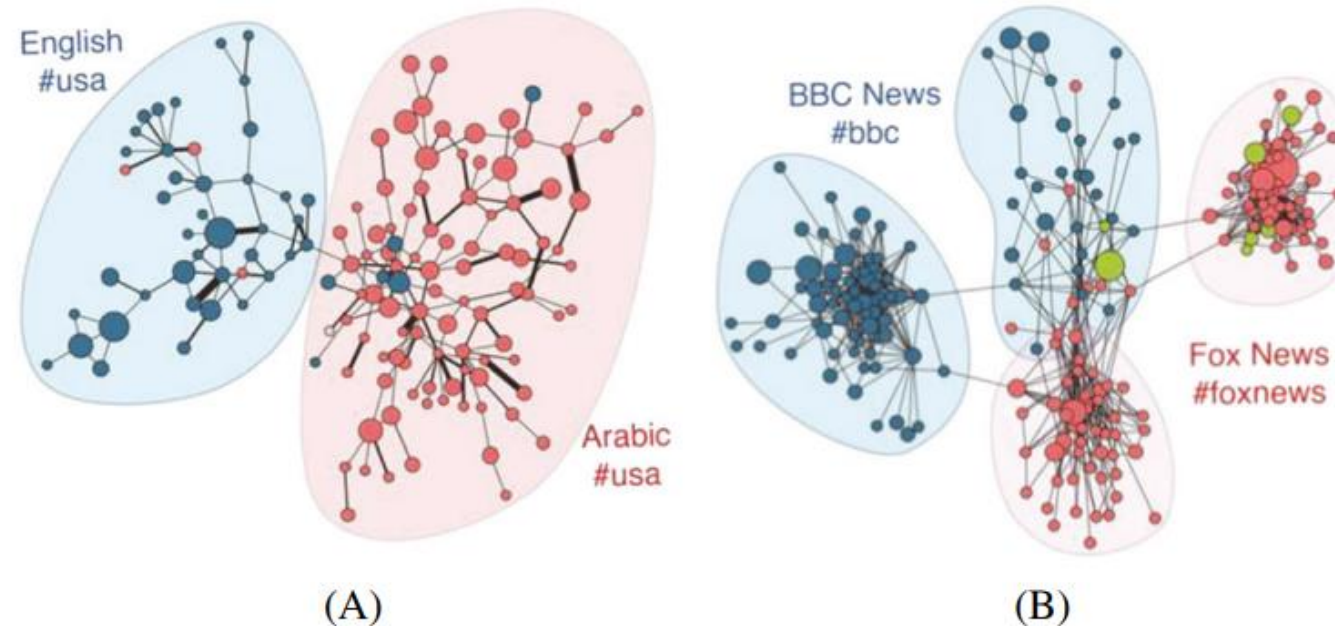


Figure 2. Example of communities in Twitter [12]. (A) Community structure based on retweets among Twitter users sharing the hashtag #USA. Blue nodes represent English users, and red nodes are Arabic users. Node size and link weight are proportional to retweet activity. (B) Community structure among Twitter users sharing the hashtags #BBC and #FoxNews. Blue nodes represent #BBC users, red nodes are #FoxNews users, and users who have used both hashtags are green. Node size is proportional to usage (tweet) activity; links represent mutual following relations.

# Assortativity and Communities (2/2)

- **Assortative communities:** in assortative networks, nodes tend to connect to other nodes with similar attributes or characteristics.
  - In this context, communities within the network often represent groups of nodes with common attributes or shared characteristics.
  - These communities are assortative because nodes within a community have similar attributes, and they are more likely to connect with each other.
- **Disassortative communities:** conversely, in disassortative networks, nodes tend to connect to nodes with dissimilar attributes. In such networks, communities can also emerge where nodes with different attributes tend to cluster together.
  - These communities are disassortative because they involve nodes with diverse characteristics that are more likely to connect within their respective communities.



# Computing Assortativity (1/3)

- Mathematically, assortativity is defined as the **Pearson Correlation Coefficient**, denoted a  $r$ .
- How assortativity is typically calculated:
  - Choose the **node attribute** you want to assess assortativity for (e.g., node degree or a specific node attribute like age or degree).
  - **For each edge** in the network, calculate the attribute values for both nodes at each end of the edge.
  - Calculate the **Pearson Correlation Coefficient** ( $r$ ) between the values of the chosen attribute for pairs of neighboring nodes.

# Computing Assortativity (2/3)

- The **formula** for the Pearson Correlation Coefficient ( $r$ ) is as follows:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

- Where:
  - $X_i$  and  $Y_i$  are the values of the two attributes for node  $i$ .
  - $\bar{X}$  and  $\bar{Y}$  are the means of the two attributes, respectively.

# Computing Assortativity (3/3)

- **Positive values** of  $r$  indicate a correlation between nodes of similar degree, while **negative values** indicate relationships between nodes of different degree.
- In general,  $r$  lies between  $-1$  and  $1$ .
  - When  $r = 1$ , the network is said to have perfect assortative mixing patterns.
  - When  $r = 0$ , the network is non-assortative (or neutral).
  - When  $r = -1$ , the network is completely disassortative.



# Some Examples

	Network	$n$	$r$	
Social networks are <b><u>assortative</u></b>	Physics coauthorship (a)	52 909	0.363	
	Biology coauthorship (a)	1 520 251	0.127	
	Mathematics coauthorship (b)	253 339	0.120	
	Film actor collaborations (c)	449 913	0.208	
	Company directors (d)	7 673	0.276	
	Internet (e)	10 697	-0.189	Biological, technological networks are <b><u>disassortative</u></b>
	World-Wide Web (f)	269 504	-0.065	
	Protein interactions (g)	2 115	-0.156	
	Neural network (h)	307	-0.163	
	Marine food web (i)	134	-0.247	
	Freshwater food web (j)	92	-0.276	

# Assortativity and Programming Languages

- NetworkX
  - <https://networkx.org/nx-guides/content/algorithms/assortativity/correlation.html#>
- Distinct possibilities:
  - Attribute assortativity
  - Numeric assortativity
  - Degree assortativity