Social Media Analytics: Behavior Analysis

Dr. Prasad Ramanathan
p_ramanathan@wilp.bits-pilani.ac.in

BITS Pilani

Pilani Campus

Acknowledgment

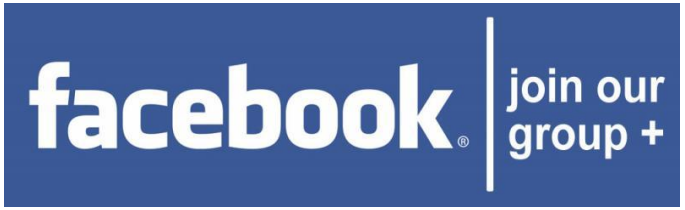
Course material from the following source is gratefully acknowledged:

R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*, Cambridge University Press, 2014.
Free book and slides at **http://socialmediamining.info/**

Examples of Behavior Analytics

innovate achieve lead

What motivates users to join an online group?



When users abandon social media sites, where do they migrate to?

Can we predict box office revenues for movies from tweets?

Behavior Analysis

To answers these questions we need to **analyze** or **predict** behaviors on social media.

Users exhibit different behaviors on social media:

- As individuals, or
- As part of a broader collective behavior.

When discussing individual behavior,

- Our focus is on one individual.

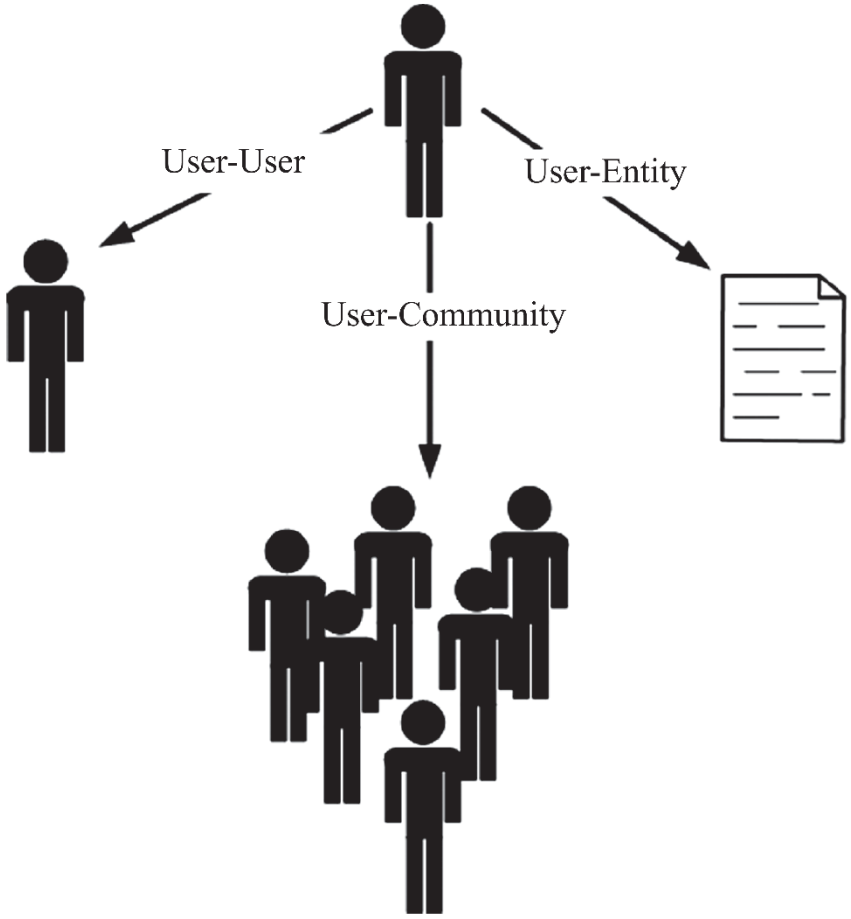
Collective behavior emerges when ***a population of individuals behave in a similar way with or without coordination or planning***.

Our Goal

To analyze, model, and predict individual and collective behavior

Individual Behavior

Types of Individual Behavior



User-User
(link generation)
befriending, sending a message, playing games, following, or inviting

User-Community
joining or leaving a community, participating in community discussions

User-Entity **(content generation)**
writing a post
posting a photo

I. Individual Behavior Analysis

Example: Community Membership in Social Media

innovate achieve lead

Why do users join communities?

- Communities can be implicit:
  - Individuals buying a product as a community, and
  - People buying the product for the first time as individuals joining the community.
- **What factors affect the community-joining behavior of individuals?**

We can observe users who join communities

- **Determine factors that are common among them**

To observe users, we require

- A population of users,
- A community C, and
- Community membership info (users who are members of C)

To distinguish between users who have already joined the community and those who are now joining it,

- We need community memberships at two times t_1 and t_2, with $t_2 > t_1$
- At t_2, we find users who are members of the community, but were not members at t_1
  - These new users form the subpopulation that is analyzed for community-joining behavior.

Community Membership in Social Media

innovate achieve lead

Hypothesis:

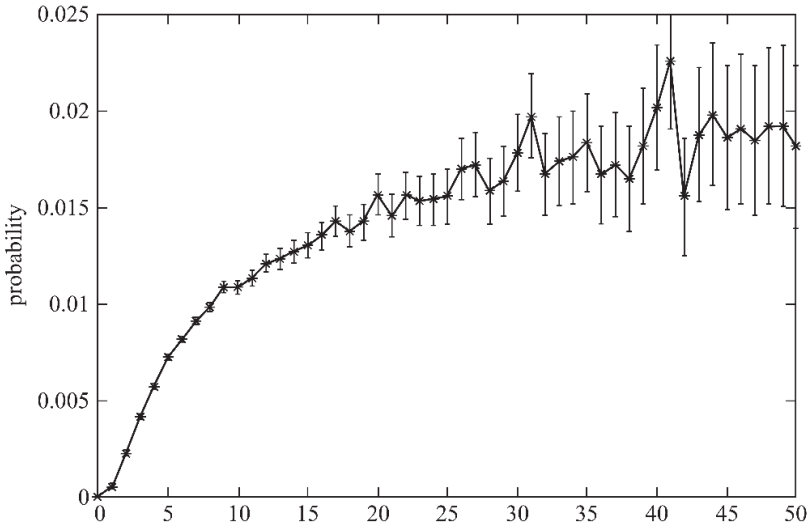
individuals are inclined toward an activity when their friends are engaged in the same activity.

A factor that plays a role in users joining a community is the number of their friends who are already members of the community.

In data mining terms,

number of friends of an individual in a community

A **feature** to predict whether the individual joins the community (i.e., **class attribute**).



Number of Friends vs Probability of Joining a Community

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006, August). Group formation in large social networks: membership, growth, and evolution. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 44-54). ACM.

Even More Features

innovate achieve lead

Feature Set	Feature
Features related to the community, C. (Edges between only members of the community are $E_C \subseteq E$.)	Number of members ($\|C\|$). Number of individuals with a friend in C (the *fringe* of C) . Number of edges with one end in the community and the other in the fringe. Number of edges with both ends in the community, $\|E_C\|$. The number of open triads: $\|\{(u,v,w)\|(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \notin E_C \wedge u \neq w\}\|$. The number of closed triads: $\|\{(u,v,w)\|(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \in E_C\}\|$. The ratio of closed to open triads. The fraction of individuals in the fringe with at least k friends in the community for $2 \leq k \leq 19$. The number of posts and responses made by members of the community. The number of members of the community with at least one post or response. The number of responses per post.
Features related to an individual u and her set S of friends in community C.	Number of friends in community ($\|S\|$). Number of adjacent pairs in S ($\|\{(u,v)\|u,v \in S \wedge (u,v) \in E_C\}\|$). Number of pairs in S connected via a path in E_C. Average distance between friends connected via a path in E_C. Number of community members reachable from S using edges in E_C. Average distance from S to reachable community members using edges in E_C. The number of posts and response made by individuals in S. The number of individuals in S with at least 1 post or response.

Feature Importance Analysis

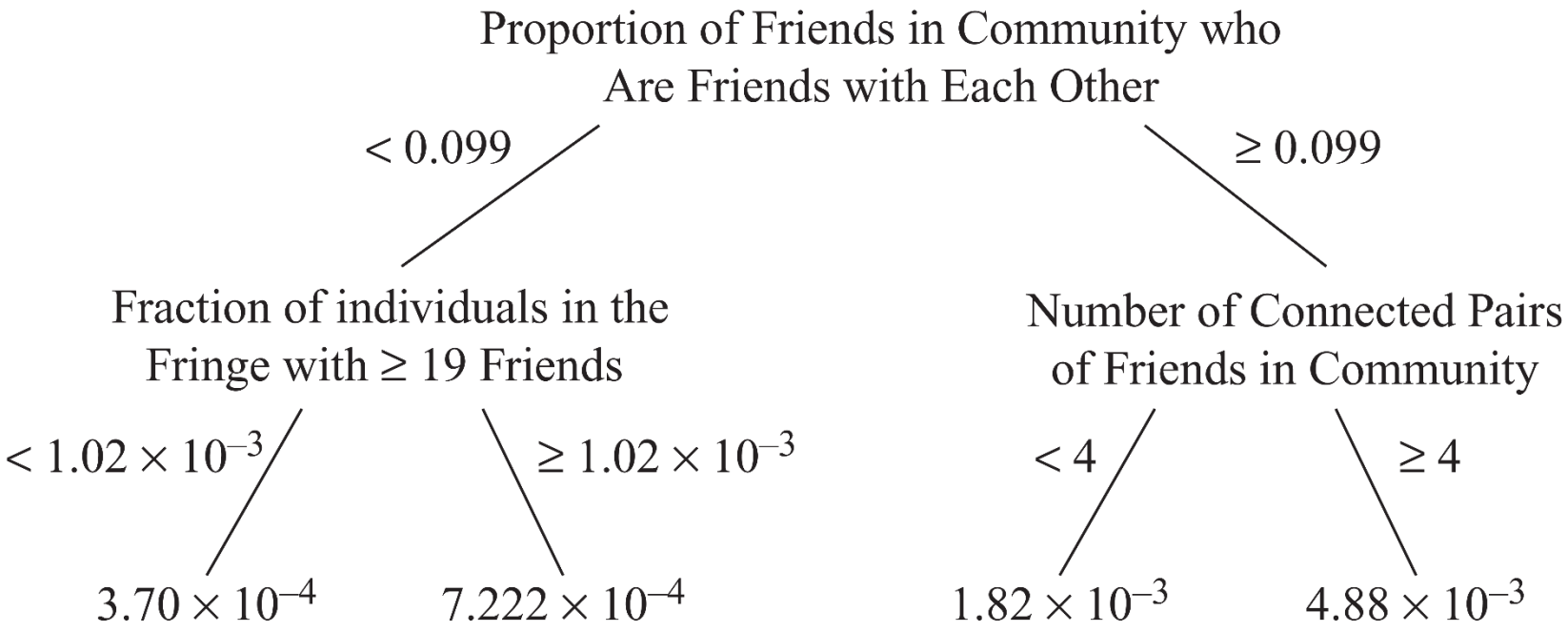
Which feature can help best determine whether individuals will join or not?

I. We can use any feature selection algorithm, *or*

II. We can use a classification algorithm, such as decision tree learning

Most important Features are **ranked higher**

Decision Tree for Joining a Community



Are these features well-designed?

We can evaluate using classification performance metrics

Behavior Analysis Methodology

An observable behavior

The behavior needs to be observable

E.g., accurately observing the joining of individuals (and possibly their joining times)

Features:

Finding data features (covariates) that may or may not affect (or be affected by) the behavior

We need a domain expert for this step

Feature-Behavior Association:

Find the relationship between features and behavior

E.g., use decision tree learning

Evaluation:

The findings are due to the features and not to externalities.

E.g., we can use

- classification accuracy
- randomization tests (discussed later!)
- or causality testing algorithms

Granger Causality

Granger Causality. Assume we are given two temporal variables $X = \{X_1, X_2, \ldots, X_t, X_{t+1}, \ldots\}$ and $Y = \{Y_1, Y_2, \ldots, Y_t, Y_{t+1}, \ldots\}$. Variable X "*Granger causes*" variable Y when historical values of X can help better predict Y than just using the historical values of Y.

Consider a linear regression model

We can predict Y_{t+1} by using either $Y_1, Y_2 \ldots Y_t$ or a combination of $X_1, X_2 \ldots X_t$ and $Y_1, Y_2 \ldots Y_t$

$$Y_{t+1} = \sum_{i=1}^{t} a_i Y_i + \epsilon_1$$

$$Y_{t+1} = \sum_{i=1}^{t} a_i Y_i + \sum_{i=1}^{t} b_i X_i + \epsilon_2$$

If $\varepsilon_2 < \varepsilon_1$ then X Granger Causes Y

Why is this not causality?

II. Individual Behavior Modeling

Individual Behavior Modeling

innovate achieve lead

Models in
Economics, Game Theory, and Network Science

We can use:

1. Threshold Models: we need to learn thresholds and weights

W_{ij} can be defined as the fraction of times user i buys a product and user j buys the same product **soon** after that

When is soon?

Similarly, thresholds can be estimated by taking into account the average number of friends who need to buy a product before user i decides to buy it.

What if friends don't buy the same products?

We can find the most similar individuals or items (similar to collaborative filtering methods)

2. Cascade Models

III. Individual Behavior Prediction

Individual Behavior Prediction

Most behaviors result in newly formed links in social media.

- It can be a link to a user, as in befriending behavior;
- A link to an entity, as in buying behavior; or
- A link to a community, as in joining behavior.

We can formulate many of these behaviors as a **link prediction** problem.

Link Prediction - Setup

Given a graph $G(V, E)$, let $e(u, v)$ denote edge between nodes u and v
- $t(e)$ denotes the time that the edge was formed

Let $G[t_1, t_2]$ represent the subgraph of G such that all edges are created between t_1 and t_2
- i.e., for all edges e in this subgraph, $t_1 < t(e) < t_2$.

Given four time stamps $t_{11} < t_{12} < t_{21} < t_{22}$ a link prediction algorithm is given
- The subgraph $G(t_{11}, t_{12})$ (**training interval**) and
- Is expected to predict edges in $G(t_{21}, t_{22})$ (**testing interval**).

We can only predict edges for nodes that exist in the **training period**

Let $G(V_{train}, E_{train})$ be our **training graph**. Then, a link prediction algorithm generates a sorted list of most probable edges in

$$V_{train} \times V_{train} - E_{train}$$

Link Prediction - Algorithms

Assign $\sigma(x, y)$ to every edge $e(x, y)$

Edges sorted by this value in decreasing order will form our ranked list of predictions

Any similarity measure between two nodes can be used for link prediction;

- Network measures (**Chapter 3**) are useful here.

Some well-known methods

- Node Neighborhood-Based Methods
- Path-Based Methods

Collective Behavior

Collective Behavior

- First Defined by sociologist Robert Park
- **Collective Behavior:** A group of individuals behaving in a similar way
- It can be planned and coordinated, but often is spontaneous and unplanned

Examples

- Individuals standing in line for a new product release
- Posting messages online to support a cause or to show support for an individual

I. Collective Behavior Analysis

Collective Behavior Analysis

- We can analyze collective behavior by analyzing individuals performing the behavior
- We can then put together the results of these analyses
  - The result would be the **expected behavior** for a large population
- **OR**, we can analyze the population as a whole
  - Not very popular for **analysis**, as individuals are ignored
  - Popular for **Prediction** purposes

Example – Analyzing User Migrations

Users migrate in social media due to their limited time and resources

- Sites are interested in keeping their users, because they are valuable assets that help contribute to their growth and generate revenue by increasing traffic

Two types of migrations:

- **Site migration**: For any user who is a member of two sites S_1 and S_2 at time t_i, and is only a member of S_2 at time $t_j > t_i$, then the user is said to have migrated from site S_1 to S_2.

- **Attention Migration:** For any user who is a member of two sites S_1 and S_2 and is active at both at time t_i, if the user becomes inactive on S_1 and remains active on S_2 at time $t_j > t_i$, then the user's attention is said to have migrated away from site S_1 and toward site S_2.

Collective Behavior Analysis - Example

Activity (or inactivity) of a user can be determined by observing the user's actions performed on the site.

We can consider a user active in $[t, t + X]$, if the user has performed at least one action on the site during this period

- Otherwise, the user is considered inactive.

The interval could be measured at different granularity levels

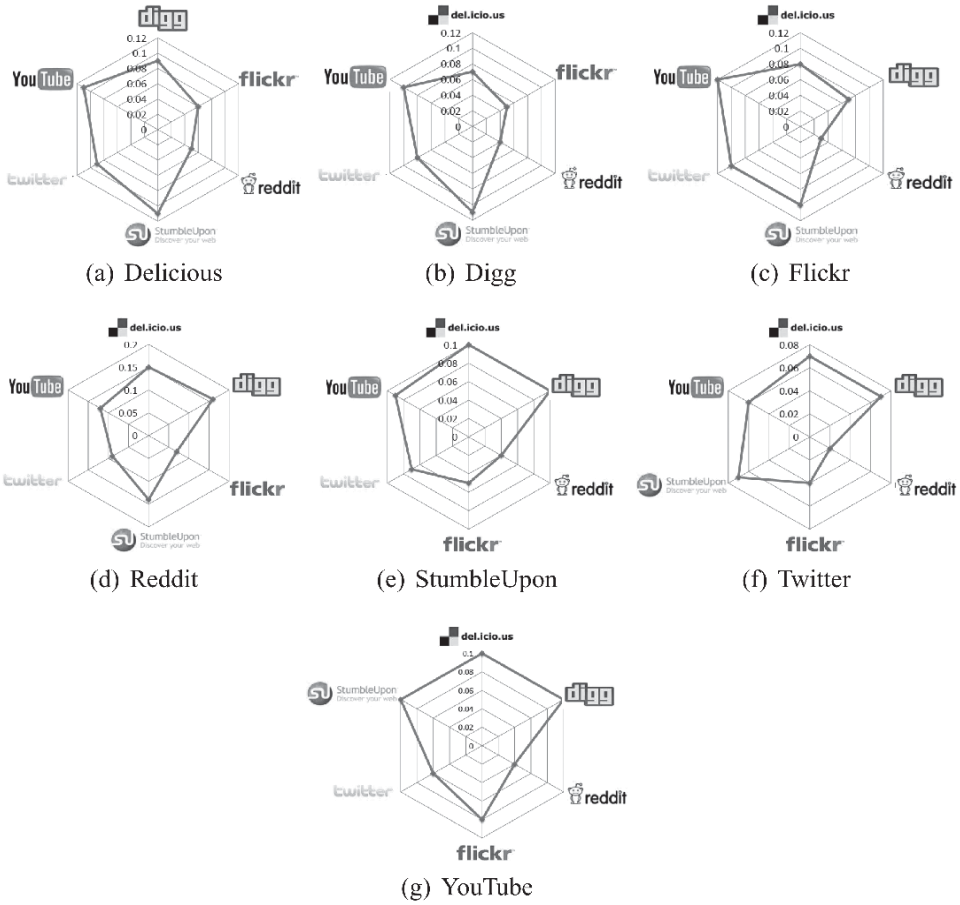
- E.g., days, weeks, months, and years.
- It is common to set $X = 1$ month.

We can analyze migrations of **individuals** and then measure the rate at which the **populations** are migrating across sites.

- We can use the methodology for individual behavior analysis

The Observable Behavior

- Sites migration is rarely observed
- Attention migration is clearly observable
- We need to take multiple steps to observe it:
  - Users are required to be identified on multiple networks (challenging!)
    - Some ideas: **John.Smith1** on Facebook is **JohnSmith** on Twitter



(a) Delicious (b) Digg (c) Flickr

(d) Reddit (e) StumbleUpon (f) Twitter

(g) YouTube

Features

innovate achieve lead

- **User Activity**: more active users are less likely to migrate
  - e.g., number of tweets, posts, or photos
- **User Network Size:** a user with more social ties (i.e., friends) in a social network is less likely to move
  - e.g., number of friends
- **User Rank:** a user with high status in a network is less likely to move to a new one where he or she must spend more time getting established.
  - e.g., centrality scores
  - External rank: your citations, how many have referred to your article, …

Feature-Behavior Association

- Given two snapshots of a network, we know if users migrated or not.
- Let vector $\mathrm{Y} \in \mathbb{R}^n$ indicate whether any of our n users have migrated or not.
- Let $X_t \in \mathbb{R}^{3\times n}$ be the features collected (activity, friends, rank) for any one of these users at time stamp t.
- The correlation between features X and labels Y can be computed via logistic regression.
- How can we verify that this correlation is not random?

Evaluation Strategy

innovate achieve lead

To verify if the correlation between features and the migration behavior is not random

- We can construct a random set of migrating users
  - compute X_{Random} and Y_{Random} for them
- Find the correlation between these random variables (e.g., regression coefficients) and it should be significantly different from what we obtained using real-world observations

We can use χ^2 (Chi-square) test for significance testing

From Original Dataset

From Random Dataset

$$\chi^2 = \sum_{i=1}^{n} \frac{(A_i - R_i)^2}{R_i}$$

II. Collective Behavior Modeling

Collective Behavior Modeling

Collective behavior can be conveniently modeled using some of the techniques discussed in **Chapter 4 - Network Models.**

We want models that can mimic characteristics observable in the population.

In network models, node properties rarely play a role

Reasonable for modeling collective behavior.

III. Collective Behavior Prediction

Collective Behavior Prediction

- From previous chapters, we could use
  - Linear Influence Model (LIM)
  - Epidemic Models
- Collective behavior can be analyzed either in terms of
  1. individuals performing the collective behavior or
  2. based on the population as a whole. (**More Common**)
- When predicting collective behavior,
  - We are interested in predicting the intensity of a phenomenon, which is due to the collective behavior of the population
  - e.g., how many of them will vote?
- We can utilize a data mining approach where features that describe the population well are used to predict a response variable
  - i.e., the intensity of the phenomenon
- A **training-testing** framework or correlation analysis is used to determine the generalization and the accuracy of the predictions.

Predicting Box Office Revenue for Movies

1. Set the target variable that is being predicted

In our example: the revenue that a movie produces.

The revenue is the direct result of the collective behavior of going to the theater to watch the movie.

2. Identify features in the population that may affect the target variable

the average hourly number of tweets related to the movie for each of the seven days prior to the movie opening (seven features)

The number of opening theaters for the movie (one feature).

3. Predict the target variable using a supervised learning approach, utilizing the features determined in step 2.

4. Measure performance using supervised learning evaluation.

The predictions using this approach are closer to reality than that of the Hollywood Stock Exchange (HSX), which is the gold standard for predicting revenues for movies

Generalizing the Idea

innovate achieve lead

- **Target variable y**
  - Some feature A that quantifies the attention
  - Some feature P that quantifies the publicity

- Train a regression model

$$y = w_1 A + w_2 P + \epsilon$$

Difficulties of Decision Making

- Which digital camera should I buy?
- Where should I spend my holiday?
- Which movie should I rent?
- Whom should I follow?
- Where should I find interesting news article?
- Which movie is the best for **our family**?

Thank you

Social Media Analytics: Data Privacy & Ethics

Dr. Prasad Ramanathan
p_ramanathan@wilp.bits-pilani.ac.in

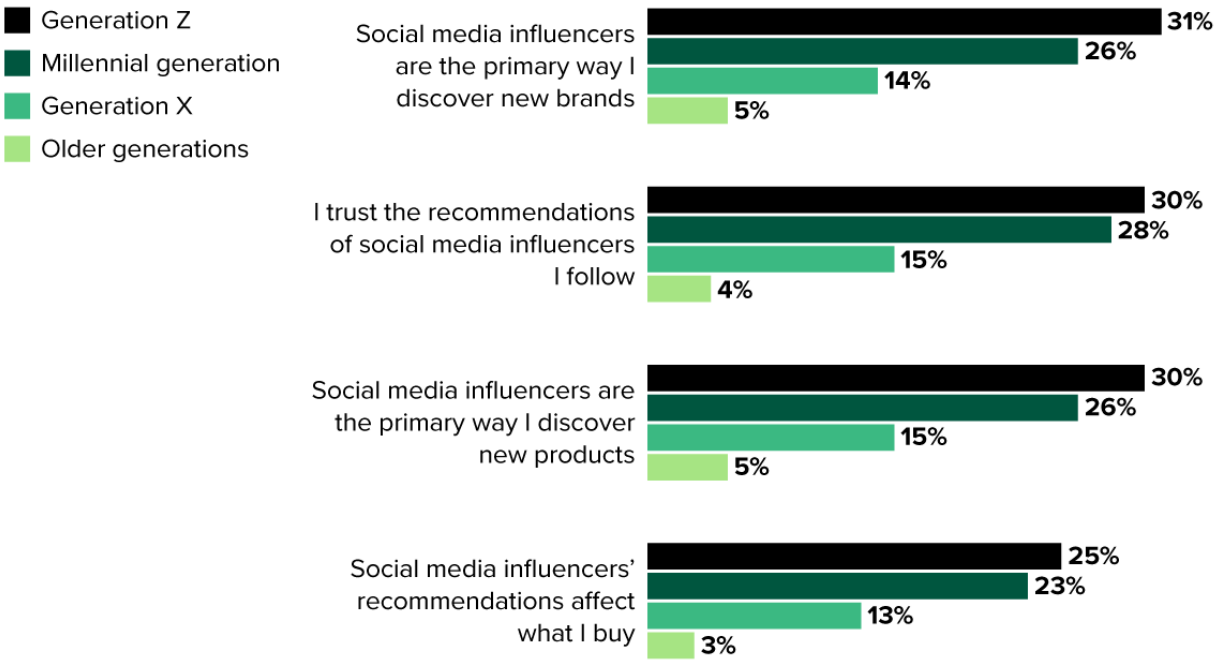
BITS Pilani

Pilani Campus

Social Media & Privacy

https://www.youtube.com/watch?v=sMLVkBxke20

"Please indicate how strongly you agree or disagree with each of the following statements about social media influencers."
(4 or 5 on a scale of 1 [strongly disagree] to 5 [strongly agree])



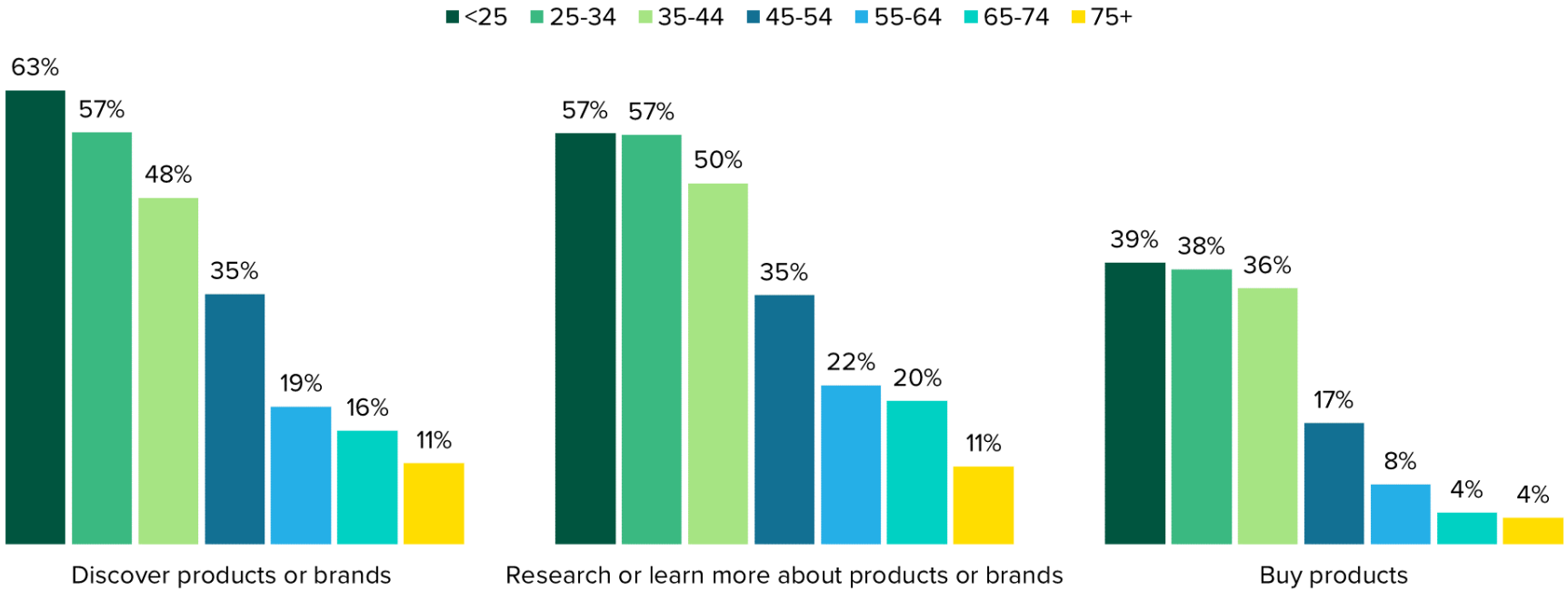
Base: 590 US online Generation Z adults (born 1997 or later); 1,467 US online Millennial adults (born 1981 to 1996); 1,163 US online Generation X adults (born 1965 to 1980); and 1,552 US online Baby Boomer and Silent Generation adults (born 1964 or earlier)

Source: Forrester Analytics Consumer Technographics® Media And Marketing Recontact Survey, 2021

Source: Forrester Research

“How frequently do you use social media to … ”

(“At least daily” and “at least weekly” responses)



Note: Sample sizes vary by age range.

Base: 269 to 963 US online adults

Source: Forrester’s Retail Topic Insights Survey, 2023

https://www.forrester.com/allSearch?activeTab=allResults&sortType=relevance&accessOnly=true&ipType=Data%20Snapshot

https://www.youtube.com/watch?v=OaLxTz1Yw7M

Ethical Dilemmas of Social Media – and How to Navigate Them

innovate achieve lead

Digital Dilemmas:

1. **Role Dilemmas** address how the person in social media can have multiple roles, creating confusion about ethical responsibilities. Such dilemmas occur when it is unclear whether a person is professionally active on a social media platform, or as a friend, client, or competitor.
2. **Tempo Dilemmas** occur because the exchanges in social media happen quickly, with an increased risk of making mistakes.
3. **Integrity Dilemmas** are concerned with how easy or difficult it is to remain committed to personal values and moral standards when representing one's organization online and being tempted or pressured to act against these.
4. **Speech Dilemmas** arise in connection with decisions about what it is acceptable to express when being active on a social media platform.
5. **Competence Dilemmas** occur when the social media experts can exploit competence gaps in their own favor, with little risk of detection. Such dilemmas occur due to the gaps in how well people understand the workings of social media.

Øyvind Kvalnes

https://www.bi.edu/research/business-review/articles/2020/07/ethical-dilemmas-of-social-media--and-how-to-navigate-them/

Ethical challenges

Facebook
- Privacy of Personal Information
- Freedom of Speech
- Data Leakage
- Identity Theft
- Fake News

Instagram
- Terms of Service & Privacy Issues
- Selling of Private Data
- Rise of Influencer Marketing

Twitter
- Fake Accounts
- Paid Tweets
- Lack of context Tweets
- Ghost Tweets
- Data Selling

Linked In
- Job Board Issues
- Erroneous information
- Lack of legal guidance
- Invasion of Privacy

Individual Level	Organizational Level
Invasion of Privacy Re-identification of Data Profiling and misuse of data Data mining risk Mis-use of free expertise & contests Anonymous Information	Competitive Pressure Poor Quality of Data Data Sharing / Sourcing Decision Making Presentation & Information

Vikas Kumar, School of Business Studies, Sharda University, Noida, India
https://orcid.org/0000-0002-6753-1557
Pooja Nanda, Sharda University, Greater Noida, India

Regulations

innovate achieve lead

General Data Protection Regulation (GDPR – EU)
The 7 data protection principles are:

1. Lawfulness, fairness, and transparency
2. Purpose limitation
3. Data minimisation
4. Accuracy
5. Storage limitations
6. Integrity and confidentiality
7. Accountability

The California Consumer Privacy Act (CCPA) – US

Digital Personal Data Protection Act (DPDP) 2023 (DPDP Act) - India

Thank you

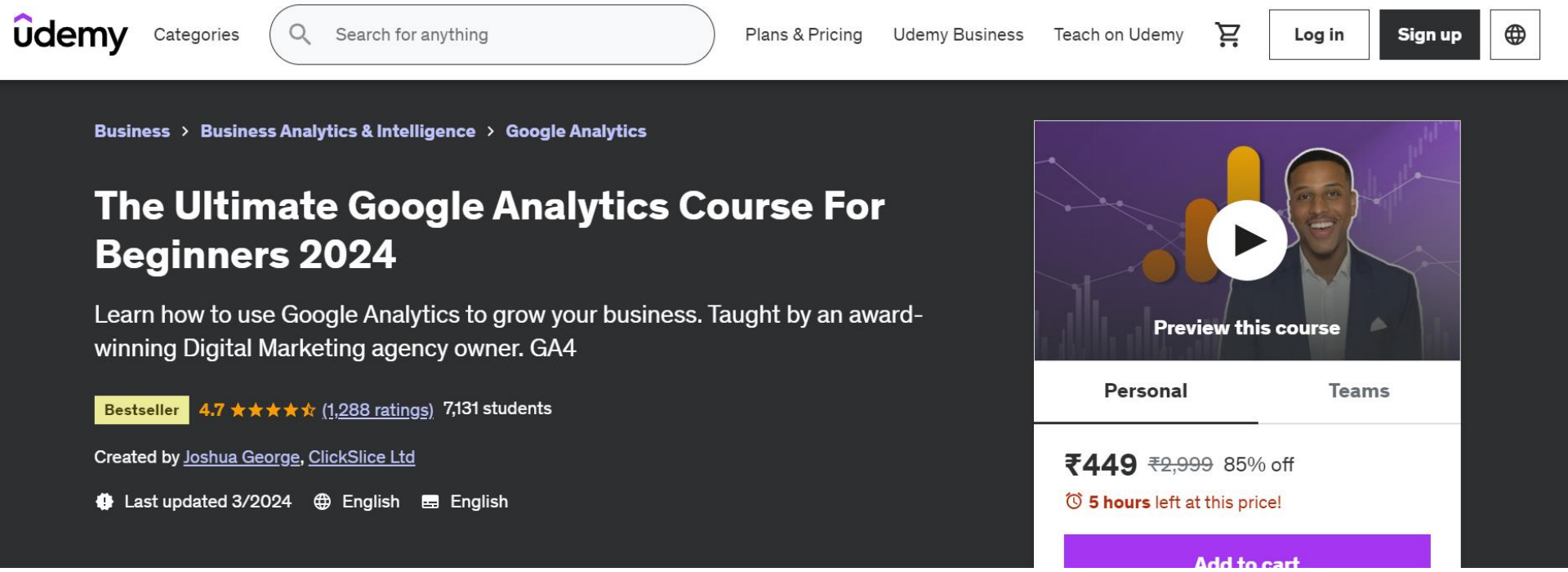
Social Media Analytics: Google Analytics GA4

Dr. Prasad Ramanathan
p_ramanathan@wilp.bits-pilani.ac.in

BITS Pilani

Pilani Campus

Reference



https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

What is Google Analytics ?

Google Analytics is a free website analytics service offered by Google that gives you insights into how users find and use your website.

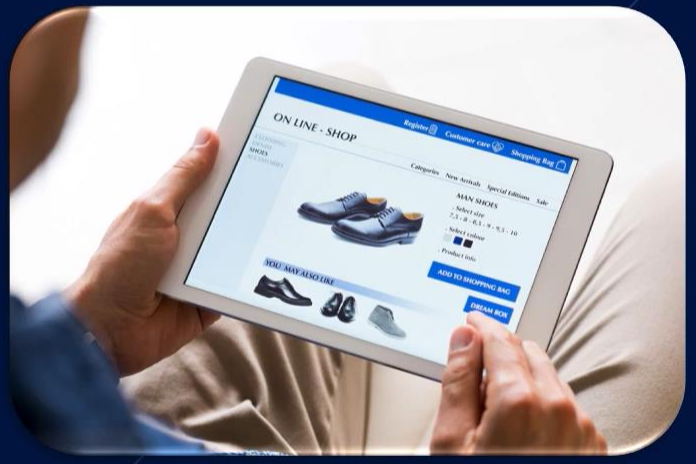
udemy

https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

Google Analytics

- Number of visitors
- Where they're coming from
- What devices they're using
- How long they spent on your site

and so much more!



https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

Make decisions based on data, not assumptions.

udemy

https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

Astonishing Facts

- Google Analytics is the most popular Web Analytics tool in the world
- It's installed on at least 10 million websites!
- It's used by 64% of the Top 500 US Retailers
- 45% of Fortune 500 companies
- 55.9% of the top 1 million domains
- All of these metrics combine to give Google Analytics a web analytics platform market share of over 82%!

https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

Google Tag Manager

innovate achieve lead

How does Google track this data?

Google Analytics tracks all of its data by a unique tracking code that YOU install on every page of your website.

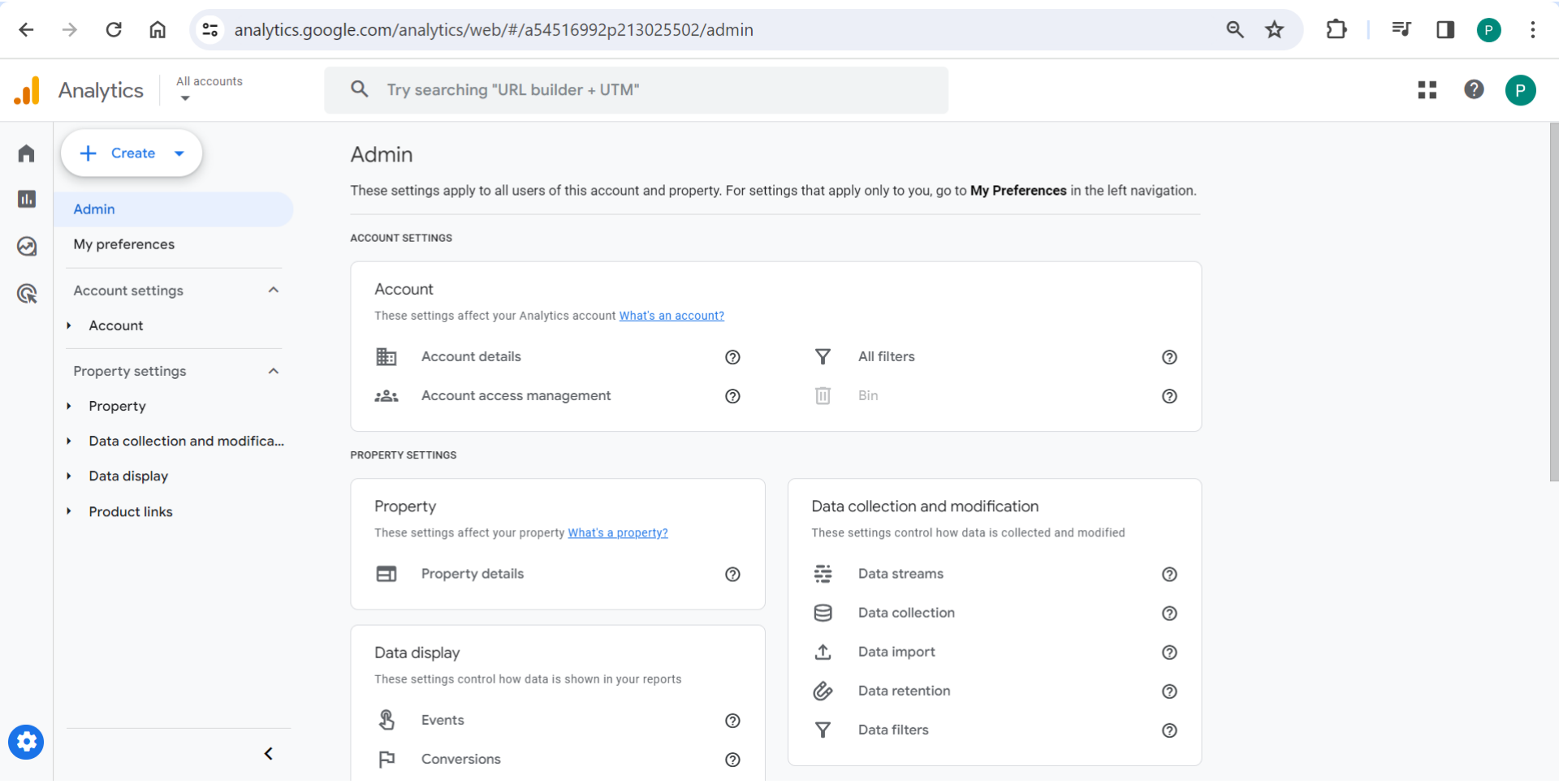
This code is essentially a small snippet of Javascript, that runs in viewers' browser when they visit those pages.

udemy

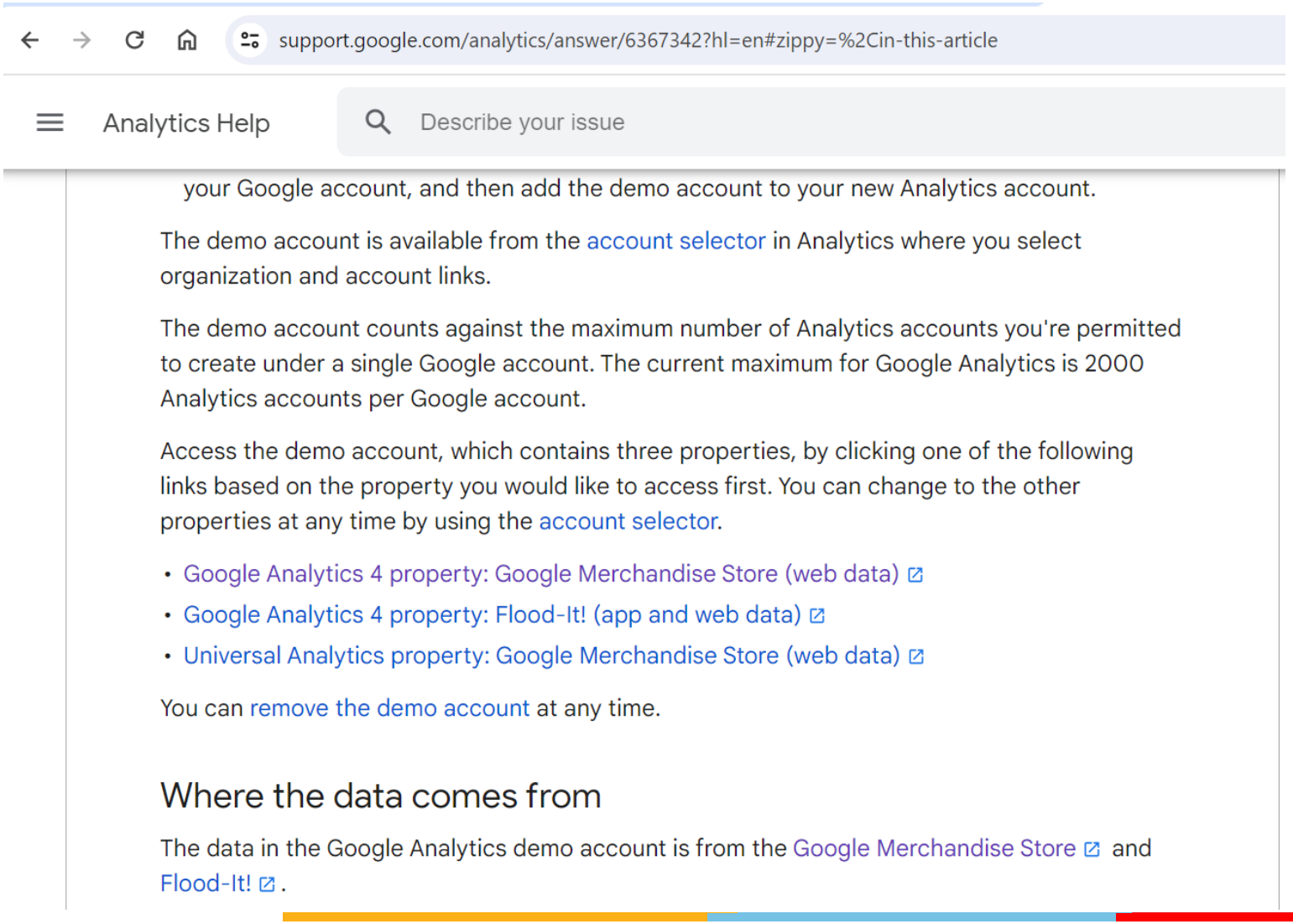
https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

https://tagmanager.google.com/#/home

Setting up Google Analytics

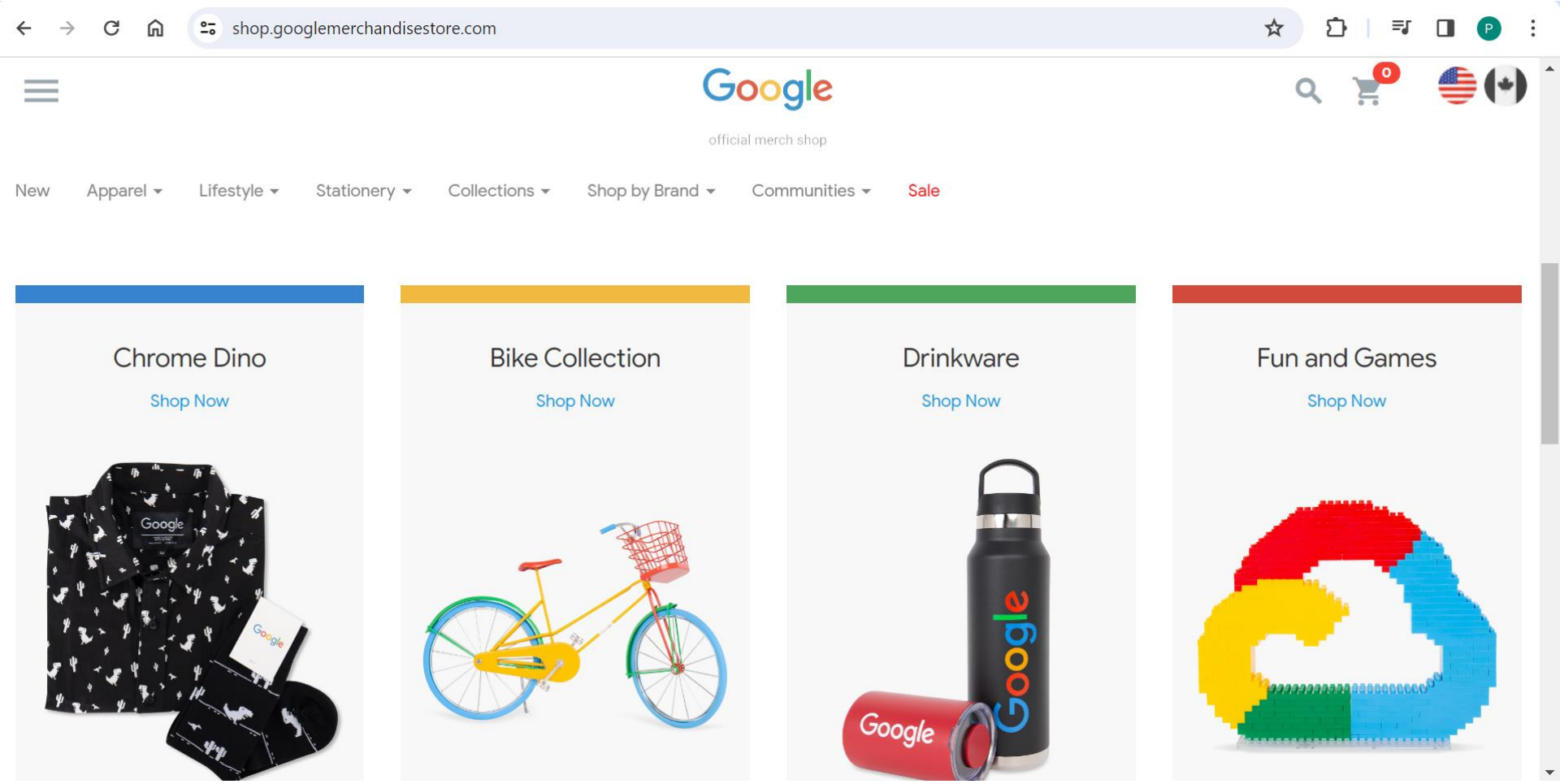


Demo account



Google Merchandise Store

https://shop.googlemerchandisestore.com/



Metrics from Google Analytics Dashboard

- How much traffic your website gets
- What country your traffic comes from
- What pages are most popular on your website
- Visitors – Demographics
- Sources & Channels that send most visitors
- Real-time reports for measuring campaign effectiveness
- Analyzing User Stickiness:
  - DAU, WAU and MAU: Daily / Weekly / Monthly Active Users
  - Ratios: DAU / MAU; DAU / WAU, WAU / MAU

Google Analytics: Glossary ...1/4

- **Automatically Collected Events** – One of the main differences between Universal Analytics and GA4 is the introduction of automatically collected events. No additional tracking is required for these events, they are automatically sent to GA by the global site tag. A full list of these events can be found here: **GA4 Automatically Collected Events**

- **Connected Site Tags** – There is a feature within GA4 called Connected Site Tags. This feature makes it possible to reuse existing Universal Analytics tagging to create a connected GA4 property. This means that you don't necessarily have to add more code to your website, in order to enable GA4 tracking – you can simply reuse existing tracking tags instead.

- **Custom Dimensions** – This is an area in GA4 that contains Custom Dimensions and Metrics. Here, you can use event parameters to create custom dimensions and metrics to be used within your reports, making labeling and understanding your data much easier.

https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

Google Analytics: Glossary …2/4

- **Data Stream** – Data Stream refers to the flow of data between your website or application and Google Analytics. Within GA4, there are three different types of data streams you can use to carry website statistics to GA – Web (for websites), Android (for Android applications), and iOS (for iOS apps).

- **Debug View** – In GA4 their Debug View allows you to test conversions and monitor events in real-time, in order to check that your tracking and reporting are working as they should. You can see events and conversions at the moment they are triggered.

- **Engaged Sessions** – This is a term used quite frequently throughout Google Analytics 4. Rather than counting all sessions within a site, GA4 focuses on engaged sessions – those where the session either lasted at least 10 seconds, had one or more conversion events, or had two or more page views. Therefore sessions in GA4 are likely to be lower than those in Universal Analytics, but they are arguably a much more valuable metric to measure.

https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

Google Analytics: Glossary …3/4

- **Enhanced Measurement** – Enhanced Measurement affects only web data streams in GA4. This feature enables a number of different events, allowing you to measure a larger number of interactions between website visitors and your content. **Click here to take a look at the full list of enhanced measurement options available within GA4.**

- **Explore** – The Explore section of GA4 is an area where users can use tables and graphs to visualize their data using highly customizable and flexible tables and graphs.

- **Life Cycle** – The Life Cycle section of GA4 contains reports that help analyze data by the stage your customers are within the overall purchase journey. Within this section, you will find reports on user acquisition, engagement, monetization, and retention.

https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

Google Analytics: Glossary …4/4

- **Monetisation** – The monetization reports in GA4 make it easier to analyze purchase activity on your website/app. This is where you'll find the data previously stored in the Conversions > E-commerce area of Universal Analytics, such as e-commerce conversion rates, product promotions, coupon uses and more.

- **Tech** – The updated Tech section of the Google Analytics profile contains data previously found within the Audience report – specifically statistics regarding user platform, operating system, app version and screen resolution. You can easily see reports mapping users by platform, OS, device and more.

- **User Snapshot** – User Snapshot is a feature of GA4, one that allows you to explore individual users and their real-time engagement with your website and/or application. Rather than just grouping all real-time data into one group, you can find out behavior and engagement data associated with individual users who are visiting your site.

https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

GA4 Sample Metrics…1/4

Acquisition – Acquisition metrics show where your traffic is originating from, be it Google searches, social media links, or other websites.

Average Session Duration – The average visit length of time a user spends on your website at any given time. This is a key metric for measuring the effectiveness and quality of your website.

Average Time on Page – The average time that users spend viewing a page or group of pages.

Bounce Rate – A bounce is a single page website visit, and so your site's bounce rate is the percentage of single page visits that your site has. Generally you want this number to be as low as possible, however sites with standalone pages such as blog articles tend to have lower bounce rates by nature.

Direct Traffic – Visitors that came directly to your site by typing your company website's URL into their browser's address bar or through a saved bookmark. Direct traffic generally indicates how many visitors already know your company and URL.

https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

GA4 Sample Metrics…2/4

Event – A ‘hit’ that tracks user interactions, such as clicks, downloads, and video plays.

Exit Page – The last page a user visits before leaving your website.

Filter – A tool that allows you to include or exclude specific data in your reports. For example, you can exclude internal company traffic so that your employees are not included in the website metrics. You can also exclude known bots.

Goal Conversion – This is the completion of an activity on your site that is important to the success of your business, such as a completed sign up for your email newsletter. You must set this up first before Google will track a goal conversion.

Landing Page – The first page that someone visits when they come to your site. Often this is the homepage.

Organic Traffic – Users who come to your website from natural (or unpaid) search engine results.

Pages/Session – The average number of pages viewed during one visit.

https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

GA4 Sample Metrics…3/4

Pageviews – The total number of website pages viewed. For example, if one user visited your homepage and the contact page, then that would count as 2 pageviews.

Referral Traffic – Visitors that landed on your website through a link on another website, such as Facebook or a site that references one of your blog articles.

Returning Visitors – Visitors that have previously visited your website (on the same device).

Search Traffic – Visitors that came to your website through a search engine such as Google or Bing.

Sessions – A session is a single continual active viewing period by a visitor. If a user visits a site several different times in one day, each unique visit counts as a session.

Source/Medium – Grouped together, source is the origin of traffic (such as bing or twitter) and medium is the category of the source (such as organic or social)

https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

GA4 Sample Metrics…4/4

Unique Visitors – The number of unduplicated visitors to your website (each user only counted once).

Unique Pageviews – Combines the pageviews from the same user in the same session, counted as one unique pageview.

Users – The number of people that have visited your site at least once during a given time period. One user could have multiple sessions, but will still be counted as a single user.

% Exit – The ratio of exits to pageviews. This indicates how often users leave page(s) compared to how many pages they view

https://www.udemy.com/course/ultimate-google-analytics-training-for-beginners/?couponCode=IND21PM

Link Google Ads → Google Analytics

1. See the full customer cycle
2. Enable auto tagging
3. Set up Remarketing campaigns

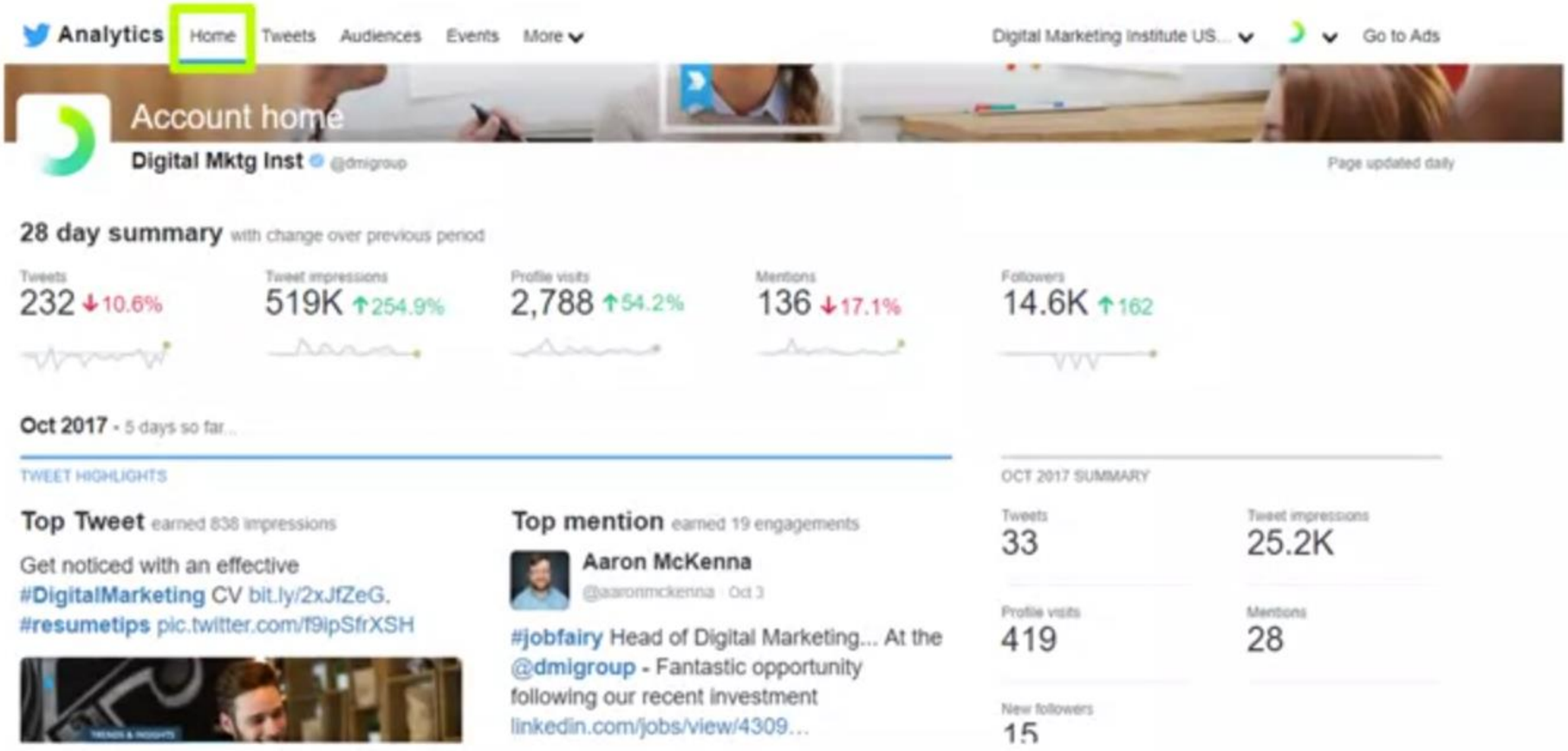
GA4 Case Studies

1. McDonald's Hong Kong
2. Claro Shop
3. Food Rescue
4. Lider

Twitter & Facebook Analytics

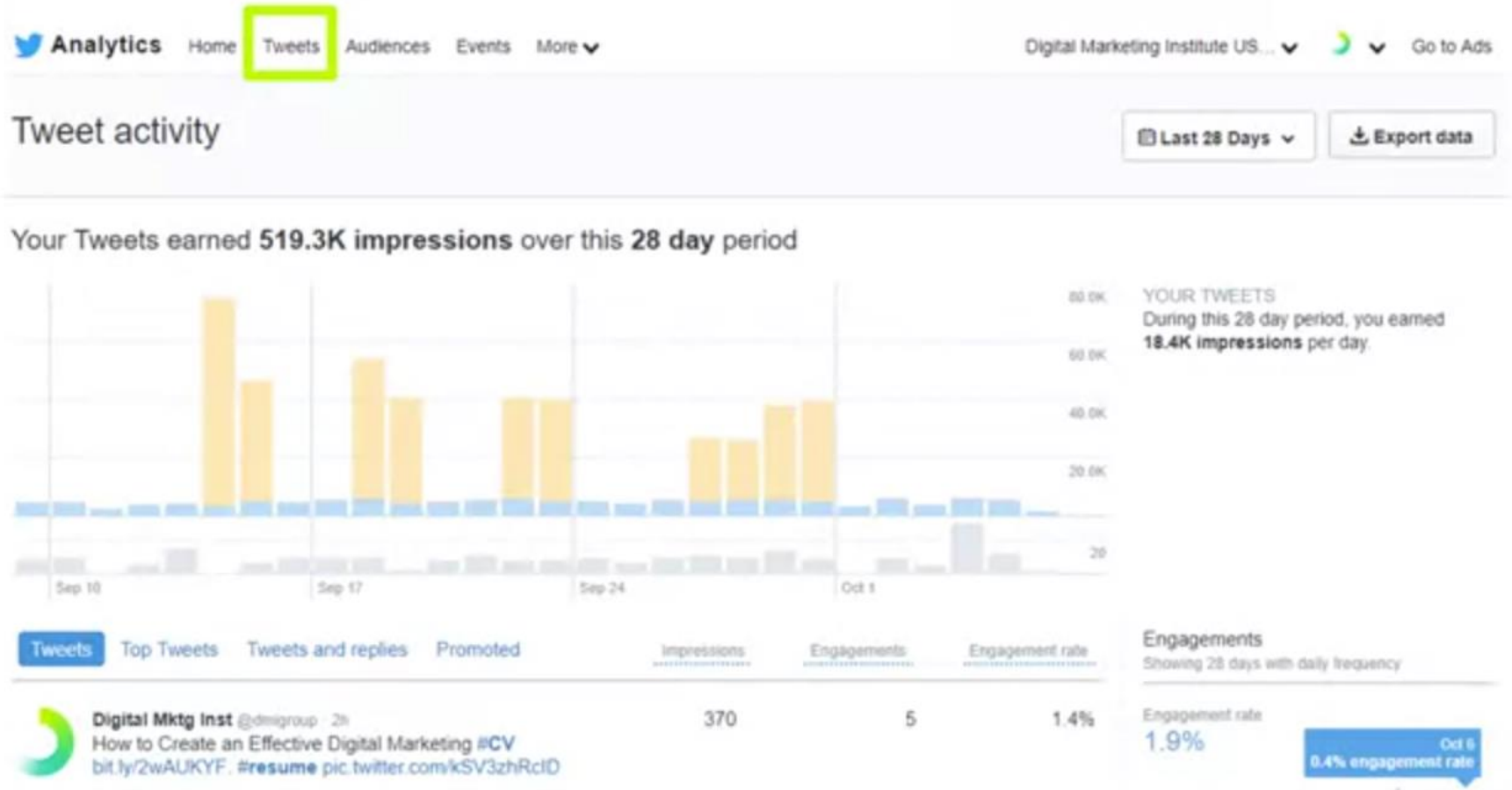
ANALYSIS AND REPORTING

Account home

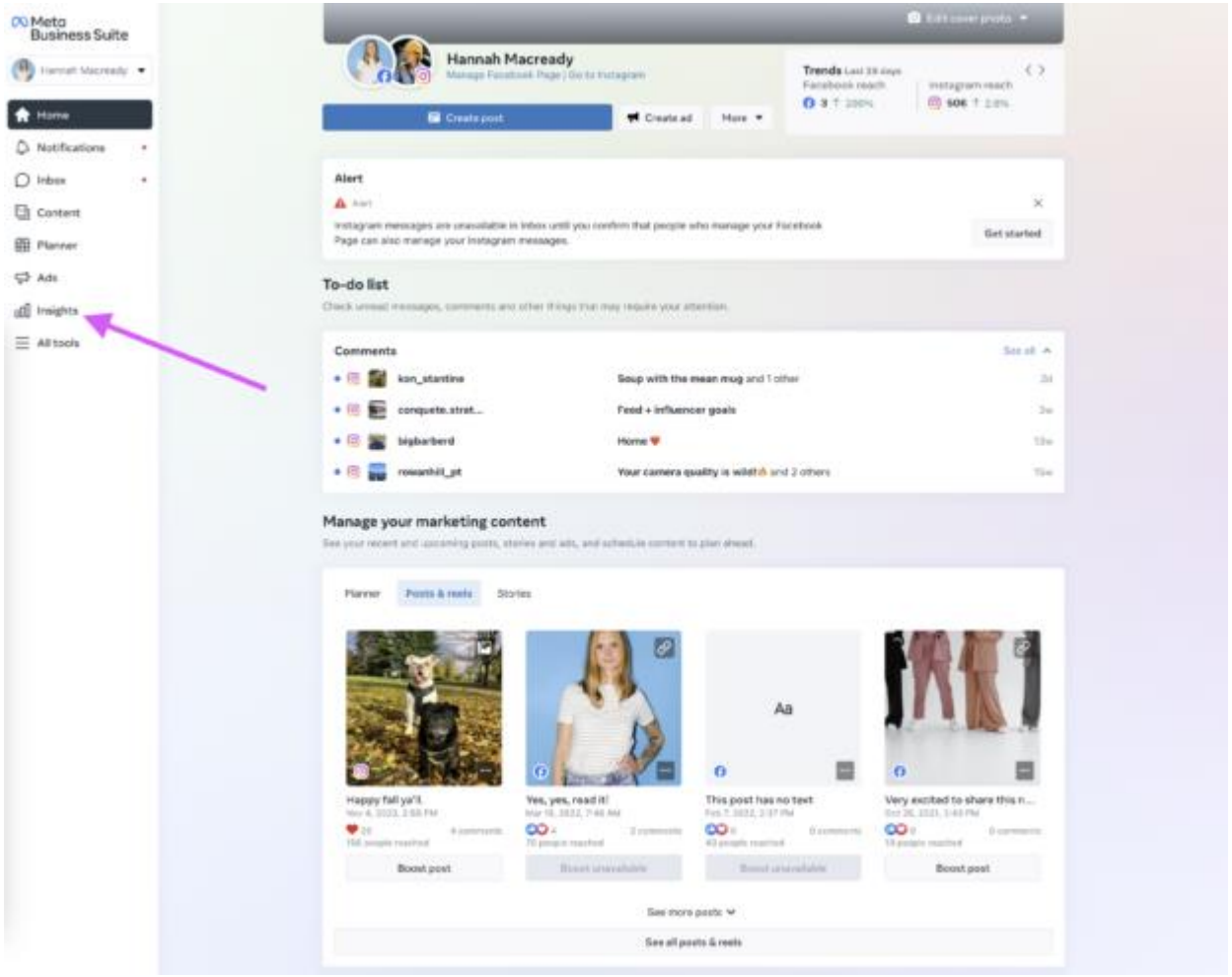


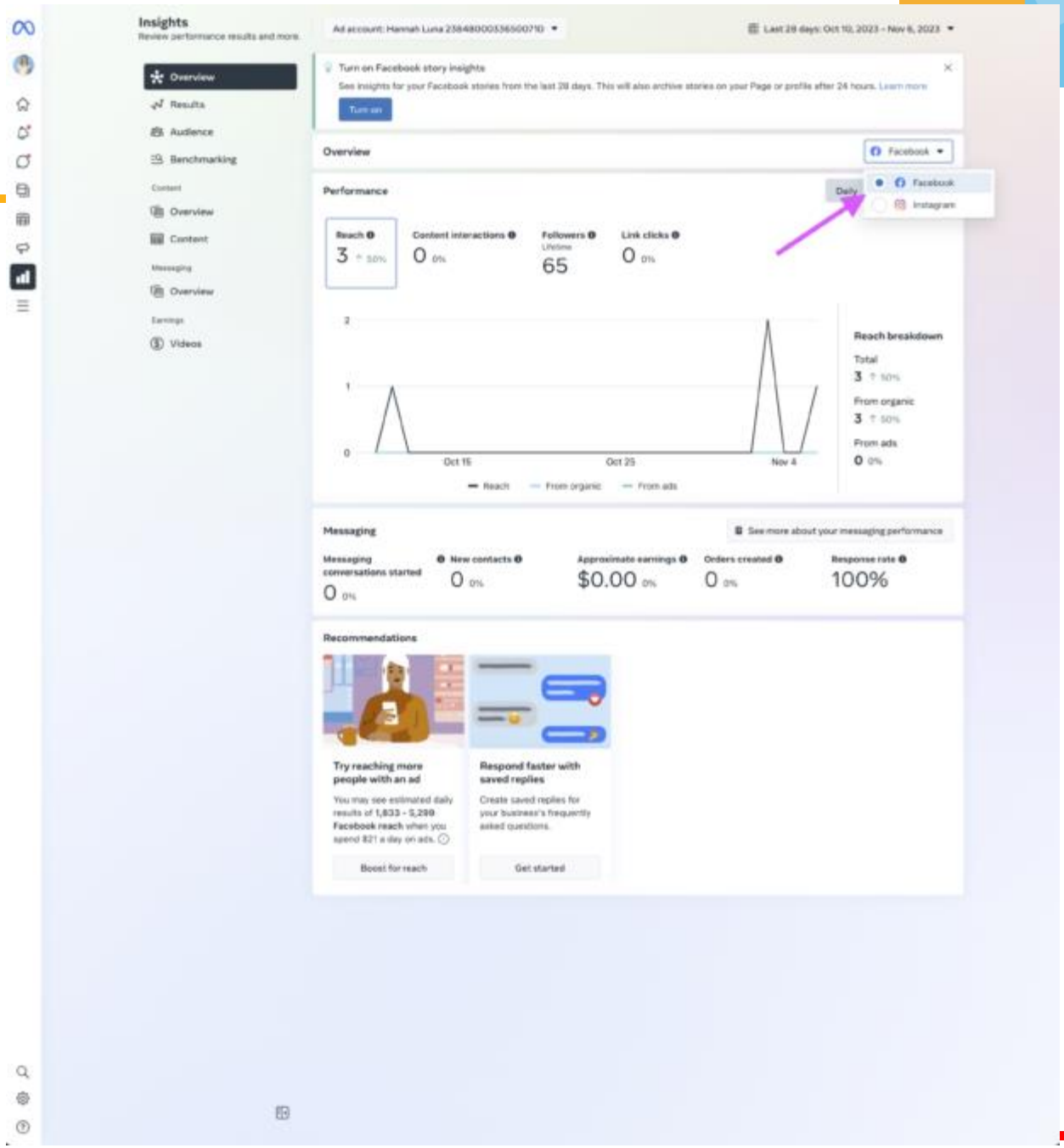
ANALYSIS AND REPORTING

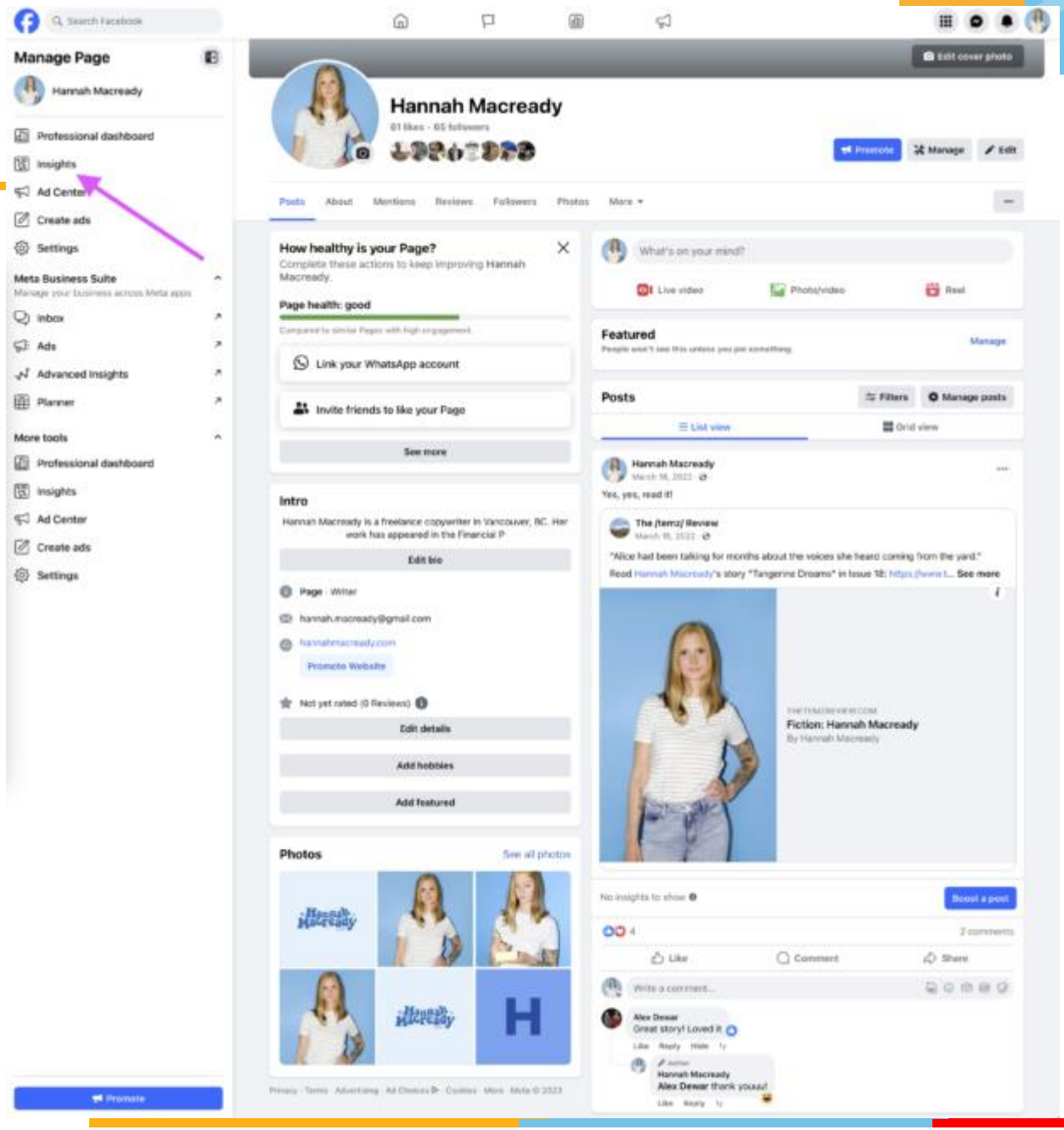
Tweet activity dashboard



Meta Business Suite




Insights
Review performance results and more.
Overview
Results
Audience
Benchmarking
Content
Overview
Content
Messaging
Overview
Earnings
Videos
Turn on Facebook story insights
Turn on
Overview
Facebook
Performance
Daily
Facebook
Instagram
Reach
3
Content interactions
0
Followers
Lifetime
65
Link clicks
0
Oct 15
Oct 25
Nov 4
Reach breakdown
Total
3
From organic
3
From ads
0
Reach
From organic
From ads
Messaging
See more about your messaging performance
Messaging conversations started
0
New contacts
0
Approximate earnings
$0.00
Orders created
0
Response rate
100%
Recommendations
Try reaching more people with an ad
Boost for reach
Respond faster with saved replies
Get started


Manage Page
Hannah Macready
Professional dashboard
Insights
Ad Center
Create ads
Settings
Meta Business Suite
Inbox
Ads
Advanced Insights
Planner
More tools
Professional dashboard
Insights
Ad Center
Create ads
Settings
Promote
Hannah Macready
Edit cover photo
Promote
Manage
Edit
Posts
About
Mentions
Reviews
Followers
Photos
More
How healthy is your Page?
Complete these actions to keep improving Hannah Macready.
Page health: good
Link your WhatsApp account
Invite friends to like your Page
See more
Intro
Edit bio
Page · Writer
hannah.macready@gmail.com
Promote Website
Not yet rated (0 Reviews)
Edit details
Add hobbies
Add featured
Photos
See all photos
What's on your mind?
Live video
Photo/video
Reel
Featured
Manage
Posts
Filters
Manage posts
List view
Grid view
Hannah Macready
Yes, yes, read it!
The /tərrz/ Review
"Alice had been talking for months about the voices she heard coming from the yard."
See more
Fiction: Hannah Macready
By Hannah Macready
No insights to show
Boost a post
Like
Comment
Share
Write a comment...
Alex Dewar
Great story! Loved it
Hannah Macready
Alex Dewar thank youuu!

Understand what's working best: See detailed insights for your accounts on Facebook and Instagram in Meta Business Suite.

Go to Meta Business Suite

Page Overview

Followers: 65

Create a post | Last 28 days

Post reach	3
Post Engagement	0
New Page likes	0
New followers	0

See Details

Content

Most recent content

Content
See content

Ads
See Ads

Audience

When you have at least 100 followers, you'll be able to see more info about your audience.

Thank you

Social Media Analytics: Application in Marketing & Other Business Operations

Dr. Prasad Ramanathan

p_ramanathan@wilp.bits-pilani.ac.in

BITS Pilani

Pilani Campus

Marketing

Marketing is the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large

Typical Steps in Marketing

1. Audience research
2. Create messages
3. Get messages in front of audience
4. Evaluate and optimize impact

Digital Marketing

- Target
- Measure

Types of Digital Marketing

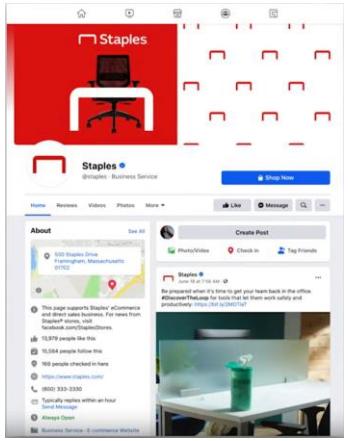
- Social Media Marketing
- Search Engine Marketing
- Search Engine Optimization
- Display Advertising
- Email Marketing
- Content Marketing

Social Media Marketing

Key Enablers in Social Media

- Connect
- Create / Share content

- Organic (free) Social Media Marketing
  - Establish business profile
  - Engage through posts
  - Connect through messaging
- Paid Social Media Marketing
  - Advertising on social media platforms



Social Media Presence

- Business
- Non-profits
  - Ice-bucket challenge to crowdsource funding for ALS

Business Accounts

- Facebook
- Instagram
- Twitter
- YouTube
- TikTok
- WhatsApp

Social Media Usage Stats

- In 2023,
  - Estimated 4.9 billion people use social media across the world
  - The most used social media platform in the world is Facebook, with 2.9 million monthly active users across the world.
  - YouTube is hot on its heels, clocking in with 2.5 million monthly active users.
  - Average person spends about 145 minutes on social media every day. Nigerian youth: 4 hrs per day; India / Phillipines – young population is more engaged
  - Short-form videos—typically less than a minute in length—capturing the attention of 66% of consumers
  - 99% use a tablet or smartphone to connect to social media, 1.32%, desktop social media users
- India
  - Facebook users 448.1 million users (i.e. 31.8% of the population)
  - Active Social Media Penetration in India is 33.4%
  - 398.0 million users who were 18 years of age or older, or 40.2 percent of the country's entire population.
  - 67.5% of all internet users (regardless of age) used at least one social networking platform.
  - Indians, on average, spend about 141.6 minutes on social media daily
  - 74.70% of internet users in India using Instagram, making it the most popular social media network there. There are 516.92 million active Instagram users in India.
  - With 492.70 million active internet users, Facebook is the second most popular platform in India, where 71.20% of internet users have profiles on the social network. Facebook is the most favored company in India among businesses, the political establishment, and the general populace, and it will continue to be a powerful influence for many years to come.
  - Twitter (42.90% penetration), LinkedIn (35.7% penetration), Moj (29.50% penetration), a short video community created locally, and Pinterest (29% penetration)

https://www.forbes.com/advisor/in/business/social-media-statistics/

Social Media Advertising Stats

- The average CTR of ads across social media was 1.21% in 2022
- 77% of businesses use social media to reach customers
- 90% of users follow at least one brand on social media
- 76% of social media users have purchased something they saw on social media: 11% buying immediately, 44% deferring online purchases for later and 21% opting to buy in-store

https://www.forbes.com/advisor/in/business/social-media-statistics/

Influencer Stats

- 50% of millennials trust influencers' product recommendations, surpassing their trust in their favorite celebrities, which stands at 38%
- 3.8 million posts on Instagram had the hashtag "ad" in 2021 (27% hike relative to 2020)
- Influencer spending hit $4.14 billion in 2022
- The minimum average cost of
  - sponsored YouTube video with 1 million views is $2,500
  - an Instagram post with 1 million followers is $1,200
  - Tik Tok post with 1 million followers is $1,034

https://www.forbes.com/advisor/in/business/social-media-statistics/

Product / Solution / Service offerings

Social Media Analytics: Key Expectations

innovate achieve lead

- Trendspotting
  - Which platforms are gaining or losing traction and popularity
  - Topics of interest that your audience is talking about (and brand mentions in conversations)
  - Types of ads that interest your audience
  - Rising influencers and products in your niche or industry
  - Types of content that your audience engages with most
- Brand Sentiment - includes all positive, neutral and negative feelings that are discussed online
  - Sentiment analysis can be used with competitor analysis because you can pinpoint new competitors and related topics your customers are buzzing about that you may have not considered before
- Value Perception – use Social Listening Tools like Google Analytics to gauge the overall customer opinion of your brand's product or service and whether it can meet their needs
- Setting Social Media Goals
  - which channels and content are performing well, so you can create actionable, realistic social media goals and objectives
- Proving ROI
  - Each time you run a new campaign, monitor your social analytics to see how the content is performing, if people are clicking over to your website and if you're generating new sales.

Reference: Sprout Social

Types of Social Media Analytics

- Performance Analysis
  - Impressions
  - Reach
  - Likes
  - Comments
  - Shares
  - Views
  - Clicks
  - Sales
- Audience Analytics
  - Age
  - Gender
  - Location
  - Device
- Competitor Analysis
  - # of followers
  - Engagement Rate
- Ad Analytics
  - Total number of active ads
  - Clicks
  - Click-through rate
  - Cost-per-click
  - Cost-per-engagement
  - Cost-per-action
  - Conversion rate
  - Total ad spend
- Influencer Analysis
  - Number of posts created per influencer
  - Total number of interactions per post
  - Audience size of each influencer
  - Hashtag usage and engagement
- Sentiment Analysis
  - Track Brand Sentiment
  - Relevant keywords and topics

Solutions from Analytics Companies: Keyhole

- Enterprise-Grade Social Listening & Analytics – for monitoring brand mentions & campaigns
- Publishing & Scheduling – For cross-platform posting
- Profile analytics – for measuring accounts' growth and engagement
- Social Media Trends – for analyzing social media trends
- Influencer Tracking – for measuring influencer campaign performance
- Historical Data – for past analytics, posts and campaigns

Reference: Keyhole

Products from Hootsuite: Social Media Analytics companies

Products

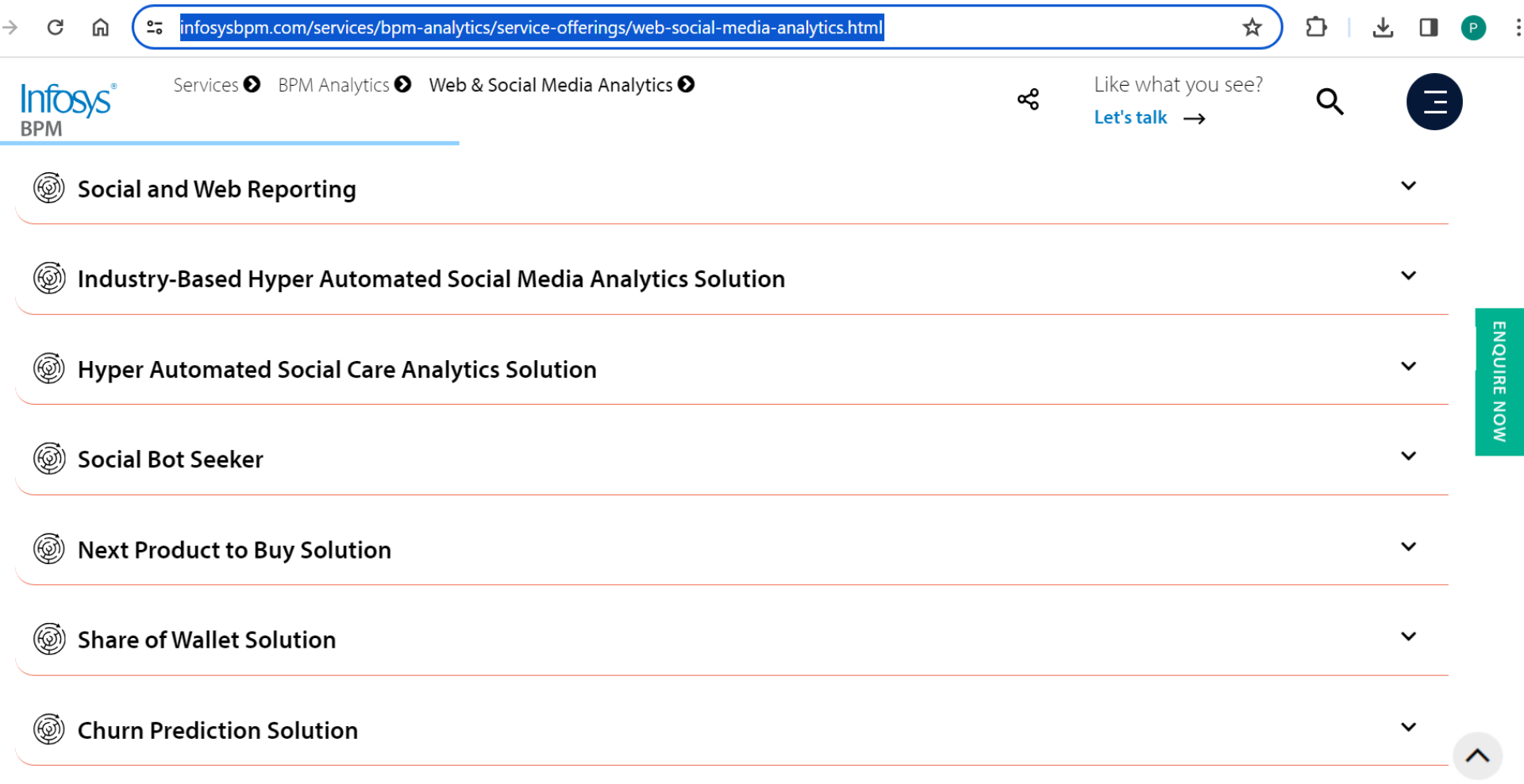
- Publish and schedule
- Engage customers
- Monitor activity
- Advertise content
- Analyze results
- Integrations

Solutions

- Customer Care
- Social Selling
- Employee advocacy
- Social Media Marketing

Reference: https://www.hootsuite.com/

Typical Service Offerings



Source: Infosys Web Site

Unleashing the potential of social media analytics for influencer marketing

Building your Social Media Marketing Strategy for 2022 and beyond

Harnessing the power of social media for marketplace management

Data privacy and ethical considerations in web and social media analytics

The role of web analytics in e-commerce: insights for online retailers

Reasons to use social media analytics

Benefits of social media analytics that you cannot deny

Incorporating web analytics into your marketing strategy

Utilising customer data analysis to boost sales

What is your share of wallet in the market?

Healthcare Data Analytics: Benefits And Use Cases

Social media analytics in 2022

A quick guide to social share of voice

Four ways enterprise analytics will evolve in 2022

Tracking and measuring web and social media analytics

Applying intelligence to analytics: The future of AI in social media

Five emerging trends shaping digital analytics

Digital analytics and its impact on digital marketing

Key benefits of digital analytics for business success

Importance of Social Network Analysis

innovate achieve lead

- **Strategic Advantage**: By understanding the connections and interactions, organisations can mobilise resources more efficiently, enhance cooperation and knowledge sharing, stimulate innovation, and gain a strategic advantage.
- **Improved Understanding**: It paves the way to understanding patterns and trends, uncovering hidden channels of information flow and decision making within and across organisations.
- **Risk Management**: SNA provides a better understanding of dependencies that could pose risks to the functionality and productivity of the system, thereby enhancing risk management.

The fantastic thing about SNA is that it **reveals the invisible** - the behind-the-scenes information flow, the influencers, gatekeepers, and liaisons. By understanding this, **businesses can enhance their strategies, communications and understand the informal and formal structures within their organisation.**

Reference: Study Smarter

Social Network Analysis Methods

- **Centrality Measures**: This gives insights into the most influential or central actors in a network.
- **Clique Analysis**: This helps identify sub-groups of nodes that are more densely connected to each other than to other nodes in the network.
- **Ego Network Analysis**: This focuses on a single node (the ego) and the nodes to which it is directly connected (the alters).
- **Cohesion Measures**: These measure how tightly knit a network is, helping you understand the strength or weakness of the overall network cohesion.

Reference: Study Smarter

Centrality Measures

- A person with a large number of friends on a social media platform would have high Degree Centrality. However, having a lot of friends doesn't necessarily mean that a person can reach others quickly, as their friends may not be well connected
- Closeness Centrality is a measure of how fast information can flow from a given node to other nodes in the network. Mathematically speaking, it is the reciprocal of the sum of the shortest paths from a node to all other nodes.
- Betweenness Centrality is a measure of the extent to which a node lies on paths between other nodes. Nodes with high betweenness centrality serve as a bridge (or a 'broker') from one part of a network to another.

Reference: Study Smarter

SNA Examples

- Understanding political structures
- Investigating the spread of diseases
- Tracing the flow of information in an organisation
- Transaction web in cryptocurrencies

Within a corporate setting:

- Insights into the informal networks that exist alongside the official organisation chart. For instance, employees often seek guidance not from their official superiors but from experienced colleagues.
- An SNA in this scenario could help to identify these individuals, measure their importance (using measures like degree centrality and betweenness centrality), and assess the impact of their eventual retirement or departure from the company.
- SNA could also showcase structural gaps where communication or collaboration is missing but necessary.

Reference: Study Smarter

SNA in Marketing

In marketing and brand strategy, SNA can help chart the landscape of social influencers. By determining the degree centrality, one can identify individuals who, due to their vast network of connections, can be instrumental in spreading content widely.

Betweenness centrality, on the other hand, can help identify those individuals who serve as critical brokers or bridges between diverse parts of the network. They might not have the highest number of connections, but they hold influence because they link different communities or groups.

A cosmetics company planning to release a new product might use SNA to identify key influencers in the beauty community. By sending products to these individuals and securing their endorsement, the company can ensure that news of the product reaches a wide audience more effectively than through traditional advertising methods.

Reference: Study Smarter

Other SNA Applications

- **Sociology**: Just as the name suggests, SNA was first developed by sociologists to understand social structures. It can unveil the complexities of human interactions, such as analysing online communities, tracking socioeconomic disparity, and studying the diffusion of cultural trends.
- **Computer Science & IT**: SNA has become a vital part of computational data analysis, primarily for the Internet and its structure. It's employed in areas like web graph analysis, cybersecurity for tracing the proliferation of malware and even optimising cloud computing networks.
- **Political Studies**: In political science, SNA is used to study policy networks, political parties, political blogs, or even to understand power structures among nations. It also aids in tracking the diffusion of political ideologies and trends.
- **Business Operations**: SNA is actively utilised to optimise organisational structures, enhance communication networks, and improve marketing strategies

Reference: Study Smarter

SNA in Business Operations

Social Network Analysis emerges as a powerful process enhancement tool. It offers a unique perspective to aid solving many business-related issues, like enhancing team collaborations, improving inter-departmental communication or even understanding customer behaviours.

Employee Interaction and Collaboration

Organisations are essentially a complex web of interactions and relationships. SNA helps to visualise this web, further enabling the organisation to understand the communication flow and thereby, promoting better collaborations. Using measures like degree centrality and betweenness centrality, one can identify key individuals who are acting as information gatekeepers.

Example: Suppose there's an individual who doesn't have an official leadership title, but their departure greatly hampers the workflow. This could possibly be because they hold a pivotal position within the informal network, answering colleagues' queries, mediating discussions, or ensuring coordination. Understanding these informal roles through SNA could significantly enhance workflow management.

Reference: Study Smarter

SNA in Business Operations

Organisational Knowledge Management

Knowledge and information in an organisation do not follow a clear-cut path as depicted by official hierarchies. Instead, it flows across organisational boundaries in rather unexpected ways. SNA allows the identification of such unconventional paths.

Example

‘T-shaped' skills, for instance, where a person has depth of knowledge in one subject (the vertical bar of the T) along with the ability to collaborate across disciplines and apply knowledge in areas of expertise other than their own (the horizontal bar of the T), are essential for innovation. SNA can help identify such individuals with 'T-shaped' skills and foster cross-disciplinary learning.

Reference: Study Smarter

SNA in Business Operations

Consumer Behaviour Analysis

On the marketing front, SNA can help understand consumer behaviours, preferences, and their decision-making process. By studying consumer networks, organisations can identify influences that impact purchasing decisions or track the diffusion of new product knowledge. With this, companies can serve more targeted advertisements and understand the potential buyer's journey.

Reference: Study Smarter

Advantages of SNA

Uncovering Hidden Relationships: The complexity of the relational data analysed is often such that making sense out of it is rather challenging. Social Network Analysis with its computational methods allows you to unravel hidden relationships and dynamics within a network, something not easily attainable through traditional data investigation techniques.

Enhanced Predictability: By determining the centrality measures (like degree, closeness, and betweenness centrality), you can predict emerging trends and behaviors within a network. In a business scenario, such predictability could enable better marketing strategies and target-specific operations.

Visualisation: One of the major advantages of SNA is its capacity to visually present complex data in an understandable form. This visual representation aids in the easy recognition of patterns, key players, and relationships.

Robustness in Various Fields: As discussed in previous sections, SNA is adept at managing multifaceted network problems in fields as diverse as business operations, sociology, computer science, and politics, among others, allowing it to adapt to a wide range of data contexts.

Reference: Study Smarter

Disadvantages of SNA

Data Entry Difficulties: The process of converting network data for SNA can be challenging and time-consuming. Collecting relational data can also prove to be more demanding than gathering simple attribute data as you need to account for connections and not just properties.

Data Privacy Concerns: With the rise of data privacy awareness, issues concerning the privacy of the network's members can pose significant challenges. Consent, purpose limitation, and data minimisation are all significant hurdles when analysing network data, especially those of a more personal nature (social networks, for example).

Interpretation Challenges: While visualisation helps represent data, SNA's interpretation is still complex due to inherent network complexity. Mistaking correlation for causality is a common issue in Social Network Analysis.

Dynamic Nature: Networks are constantly evolving and changing over time. Capturing a snapshot of the network at a single time point may therefore not provide an accurate representation as the state of the network could shift rapidly.

Reference: Study Smarter

SNA in Supply Chain Management

innovate achieve lead

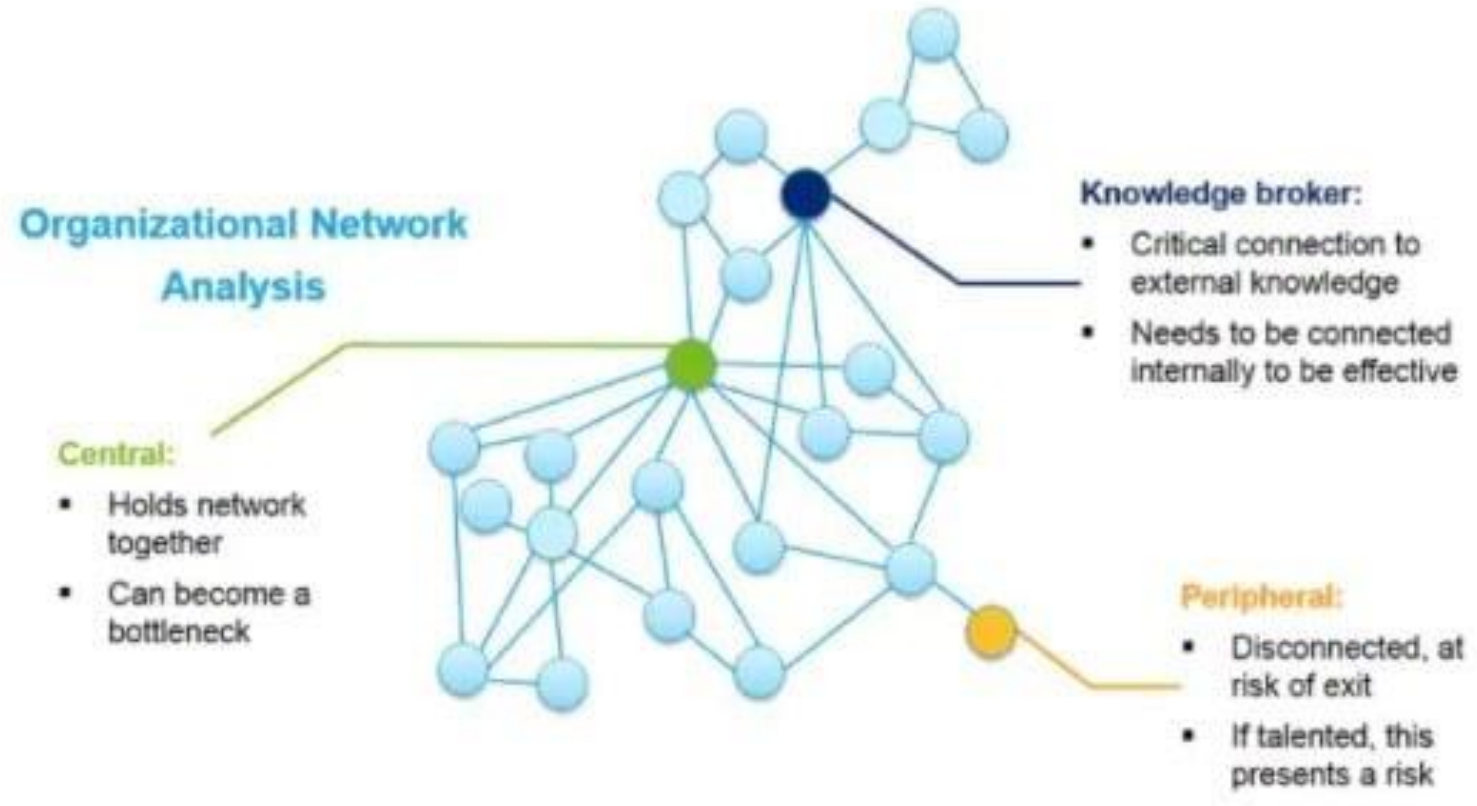
- A supply chain can be modeled into a network of supplier/consumer relations. Network analysis on the supply chain helps us improve the operation efficiency by identifying and eliminating less important nodes (suppliers/warehouses). It can help identify crucial nodes in the network and create a standby in crises or emergencies.

- Nodes include Retailers, Suppliers, Warehouses, Transporters, Regulatory agencies.

- SNA applications can help manufacturers identify more operationally critical nodes and identify potential sources to increase the number of connections to suppliers. This can also help identify any bottlenecks in the supply process and inventory management.

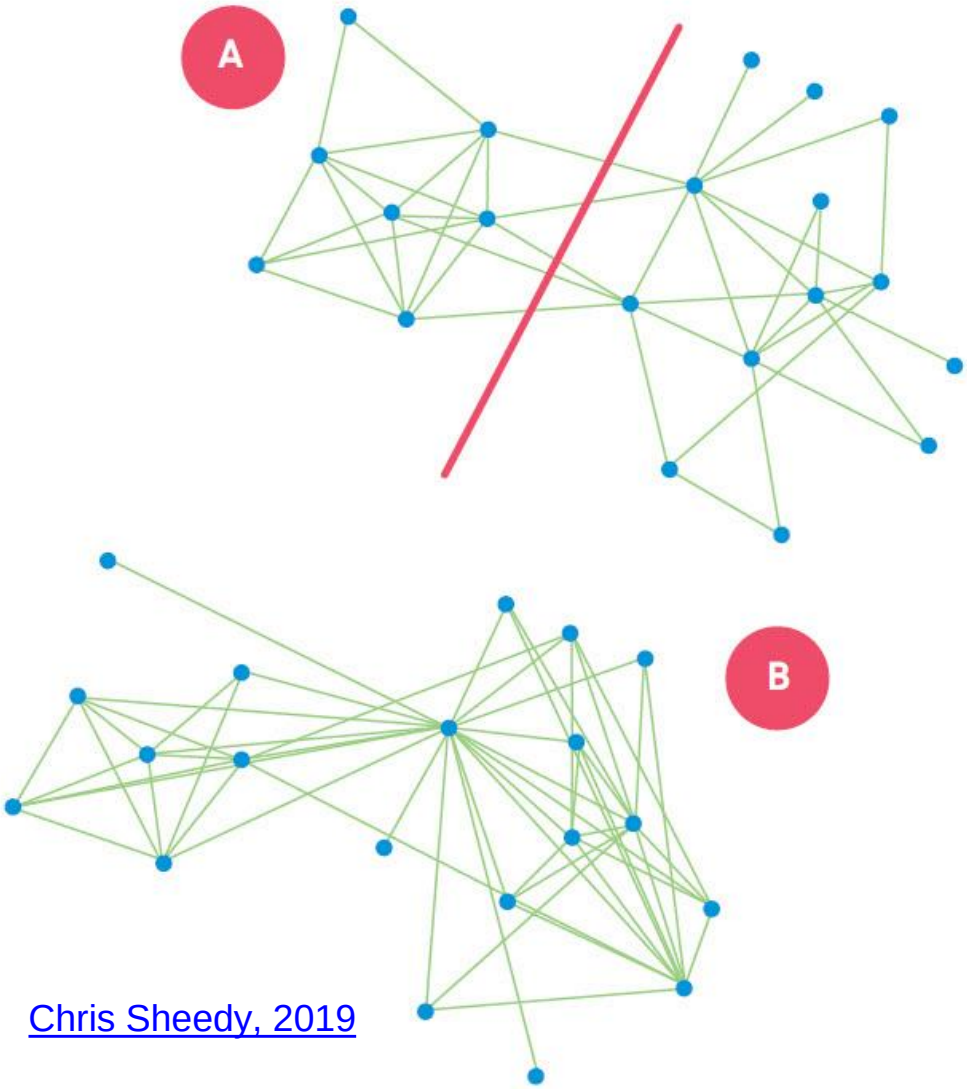
Source: https://www.latentview.com/blog/a-guide-to-social-network-analysis-and-its-use-cases/

Organizational Network Analysis



Source: https://www2.deloitte.com/us/en/pages/human-capital/articles/organizational-network-analysis.html

SNA in Teams



- Scenario (A), the group was split into two subgroups that were relatively comfortable communicating with each other.
- The group had, in fact, been two teams, brought together under a single manager. They were co-located, but still largely working as two separate groups.
- The company's management thought team building might help bring the subgroups together.
- Scenario (B) shows the team three months later. There were more connections within the group, with one individual particularly pivotal in unifying the team.
- What caused this change? Not traditional team-building exercises, but "targeted self-disclosure exercises".

Chris Sheedy, 2019

Building Trust Through Skillful Self-Disclosure

SNA in Medicine

Day nurse 1 > Night nurse 1	never gave info (0)
Day nurse 1 > Night nurse 2	often gave info (3)
Day nurse 1 > Night nurse 3	constantly gave info (4)
etc. for all night nurses	
Day nurse 2 > Night nurse 1	seldom gave info (1)
Day nurse 2 > Night nurse 2	seldom gave info (1)
Day nurse 2 > Night nurse 3	constantly gave info (4)
etc. for all night nurses	

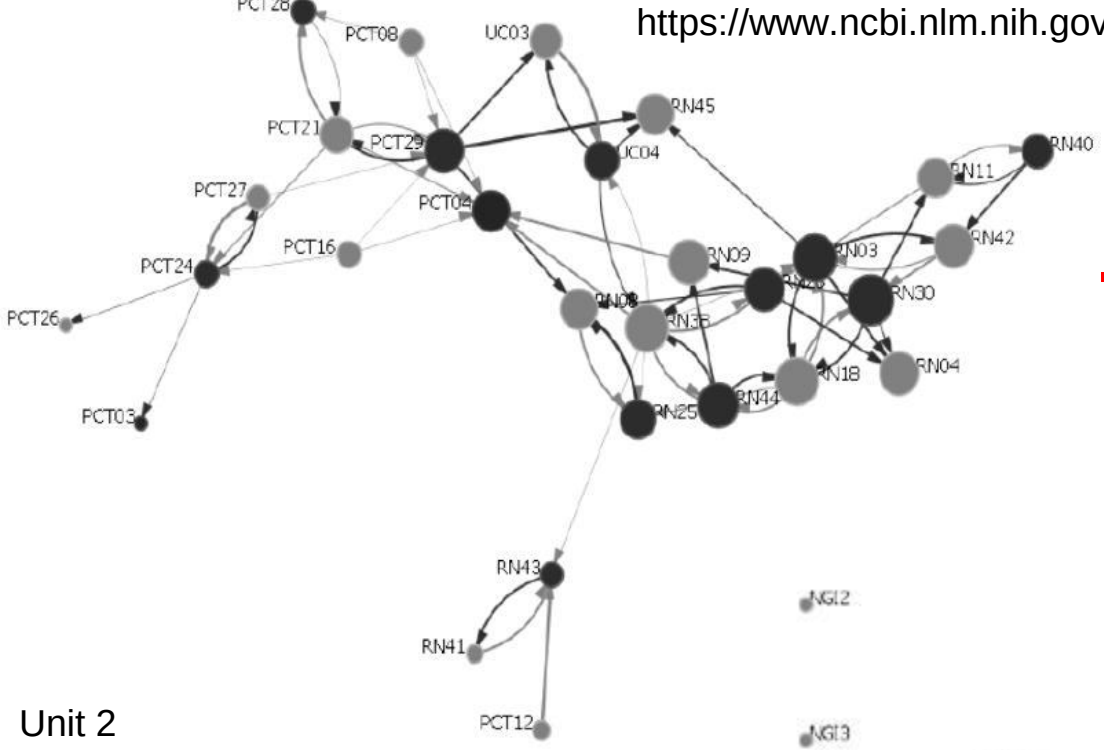
then etc. for all day nurses

https://www.statistics.com/social-network-analysis-sna-in-medicine/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4156878/

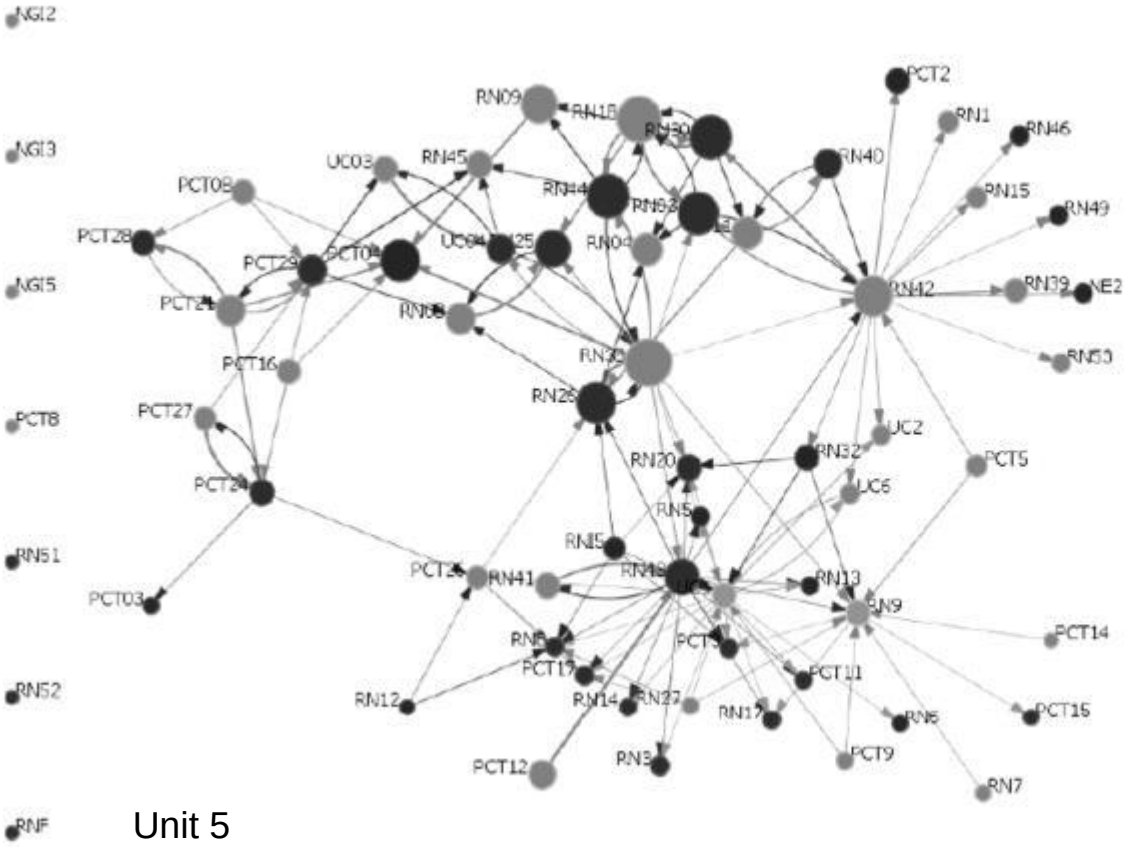
innovate achieve lead

- PCT: Patient Care Technician
- RN: Registered Nurse
- Gray circles: Day shift
- Dark circles: Night shift
- Thicker links: Higher comm frequency
- Larger node size represents higher Eigenvector Centrality values



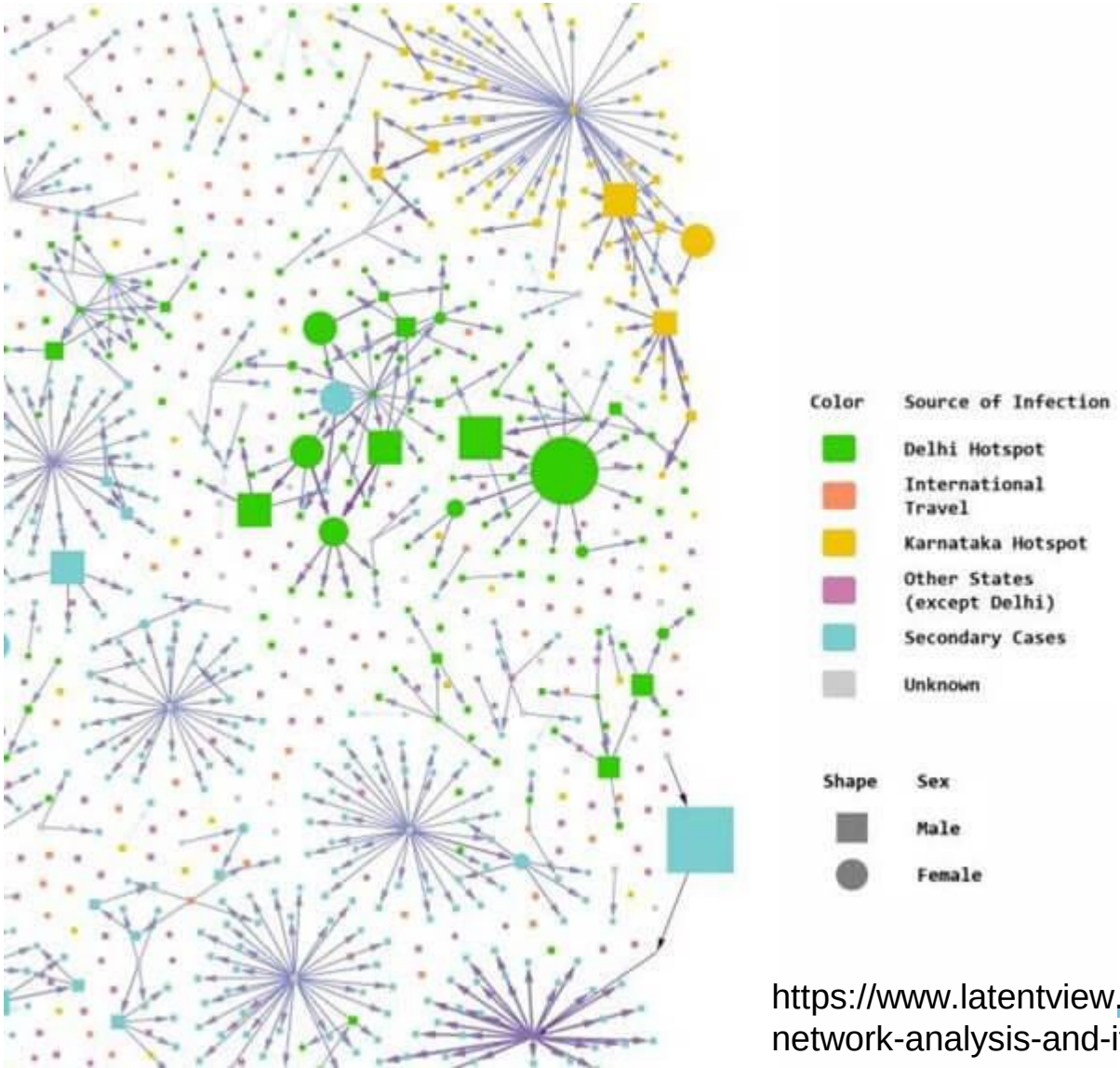
Unit 2

- Unit 2's diffusion metric is nearly twice as high (.43) as that for the much larger Unit 5 (.28). Information might diffuse quickly across the network of Unit 2
- The average Eigenvector Centrality value for Unit 2 is .45, compared with that of Unit 5 (.20). This suggests that on the smaller unit, more individuals are connected to highly connected staff.
- In Unit 5, two of the more influential RNs (high Eigenvector Centrality) are not communicating with staff not on their shift, and there are a number of "pendants" (people with single links). The pendants are usually PCTs.



Unit 5

SNA for Infectious Disease Tracking



https://www.latentview.com/blog/a-guide-to-social-network-analysis-and-its-use-cases/

Diffusion of microfinance

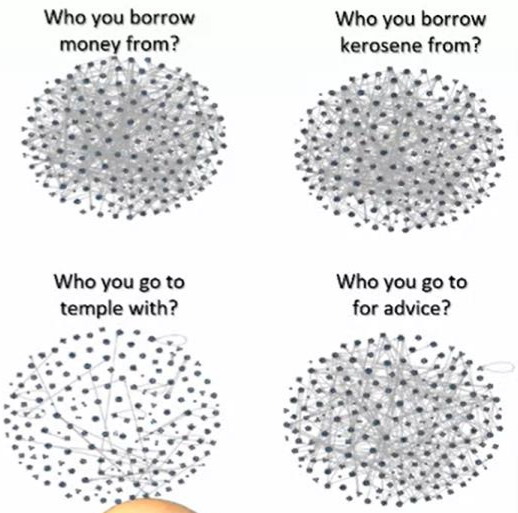
Centrality Application

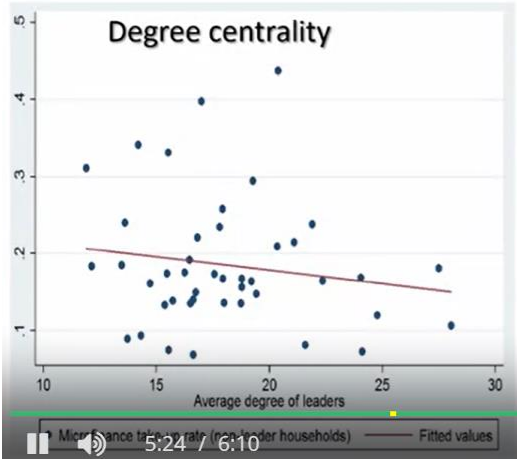
➢ **Diffusion of microfinance**

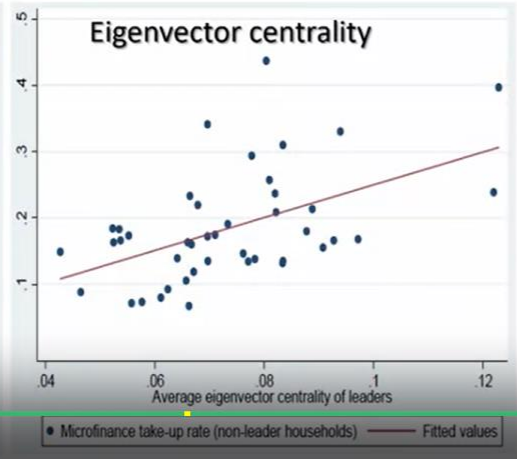
Banerjee, Chandrasekhar, Duflo & Jackson (2013). The Diffusion of Microfinance. *Science*, *341*(6144), 1236498. https://doi.org/10.1126/science.1236498

- 75 rural villages in Karnataka/ India, without microfinance
- Bank entered 43 of them and offered microfinance

- Question:
  Who to contact first to spread the innovation?
- Challenge:
  How to map the network: "who would you borrow from?"
  …they created 13 different/ multiplex networks…



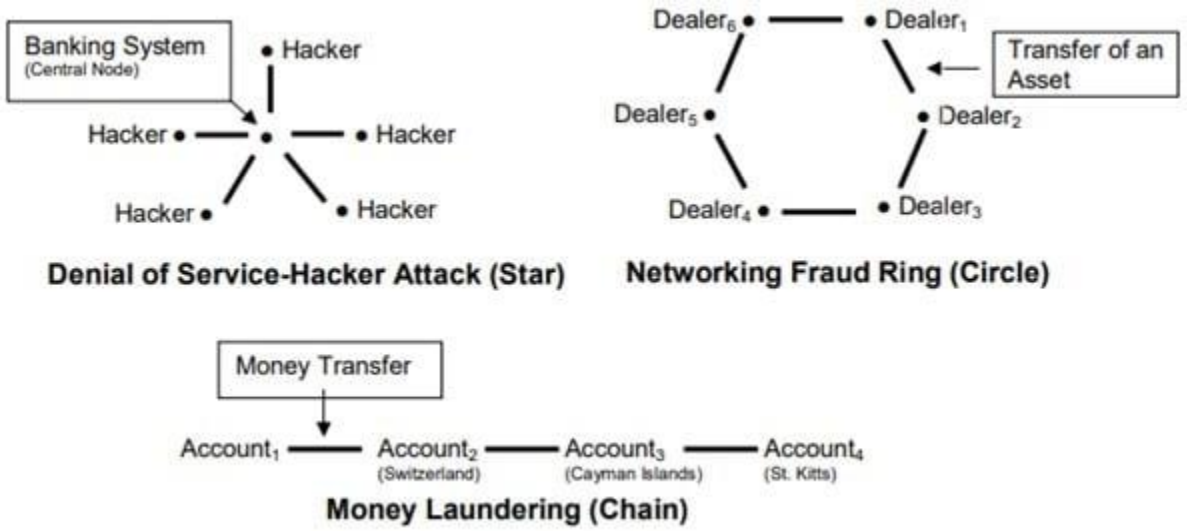




=> contact those whose friends have many friends!

Reference: https://www.coursera.org/learn/social-network-analysis/home/week/2

SNA in Finance / Fraud Detection



https://www.latentview.com/blog/a-guide-to-social-network-analysis-and-its-use-cases/

Thank you

Social Media Analytics: Application in Marketing & Other Business Operations

Dr. Prasad Ramanathan

p_ramanathan@wilp.bits-pilani.ac.in

BITS Pilani

Pilani Campus

Marketing

Marketing is the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large

Typical Steps in Marketing

1. Audience research
2. Create messages
3. Get messages in front of audience
4. Evaluate and optimize impact

Digital Marketing

- Target
- Measure

Types of Digital Marketing

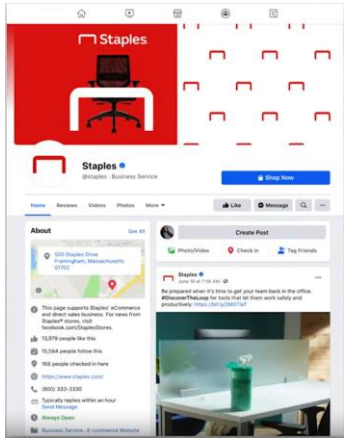
- Social Media Marketing
- Search Engine Marketing
- Search Engine Optimization
- Display Advertising
- Email Marketing
- Content Marketing

Social Media Marketing

Key Enablers in Social Media

- Connect
- Create / Share content

- Organic (free) Social Media Marketing
  - Establish business profile
  - Engage through posts
  - Connect through messaging
- Paid Social Media Marketing
  - Advertising on social media platforms



Social Media Presence

- Business
- Non-profits
  - Ice-bucket challenge to crowdsource funding for ALS

Business Accounts

- Facebook
- Instagram
- Twitter
- YouTube
- TikTok
- WhatsApp

Social Media Usage Stats

- In 2023,
  - Estimated 4.9 billion people use social media across the world
  - The most used social media platform in the world is Facebook, with 2.9 million monthly active users across the world.
  - YouTube is hot on its heels, clocking in with 2.5 million monthly active users.
  - Average person spends about 145 minutes on social media every day. Nigerian youth: 4 hrs per day; India / Phillipines – young population is more engaged
  - Short-form videos—typically less than a minute in length—capturing the attention of 66% of consumers
  - 99% use a tablet or smartphone to connect to social media, 1.32%, desktop social media users
- India
  - Facebook users 448.1 million users (i.e. 31.8% of the population)
  - Active Social Media Penetration in India is 33.4%
  - 398.0 million users who were 18 years of age or older, or 40.2 percent of the country's entire population.
  - 67.5% of all internet users (regardless of age) used at least one social networking platform.
  - Indians, on average, spend about 141.6 minutes on social media daily
  - 74.70% of internet users in India using Instagram, making it the most popular social media network there. There are 516.92 million active Instagram users in India.
  - With 492.70 million active internet users, Facebook is the second most popular platform in India, where 71.20% of internet users have profiles on the social network. Facebook is the most favored company in India among businesses, the political establishment, and the general populace, and it will continue to be a powerful influence for many years to come.
  - Twitter (42.90% penetration), LinkedIn (35.7% penetration), Moj (29.50% penetration), a short video community created locally, and Pinterest (29% penetration)

https://www.forbes.com/advisor/in/business/social-media-statistics/

Social Media Advertising Stats

- The average CTR of ads across social media was 1.21% in 2022
- 77% of businesses use social media to reach customers
- 90% of users follow at least one brand on social media
- 76% of social media users have purchased something they saw on social media: 11% buying immediately, 44% deferring online purchases for later and 21% opting to buy in-store

https://www.forbes.com/advisor/in/business/social-media-statistics/

Influencer Stats

- 50% of millennials trust influencers' product recommendations, surpassing their trust in their favorite celebrities, which stands at 38%
- 3.8 million posts on Instagram had the hashtag "ad" in 2021 (27% hike relative to 2020)
- Influencer spending hit $4.14 billion in 2022
- The minimum average cost of
  - sponsored YouTube video with 1 million views is $2,500
  - an Instagram post with 1 million followers is $1,200
  - Tik Tok post with 1 million followers is $1,034

https://www.forbes.com/advisor/in/business/social-media-statistics/

Product / Solution / Service offerings

Social Media Analytics: Key Expectations

innovate achieve lead

- Trendspotting
  - Which platforms are gaining or losing traction and popularity
  - Topics of interest that your audience is talking about (and brand mentions in conversations)
  - Types of ads that interest your audience
  - Rising influencers and products in your niche or industry
  - Types of content that your audience engages with most
- Brand Sentiment - includes all positive, neutral and negative feelings that are discussed online
  - Sentiment analysis can be used with competitor analysis because you can pinpoint new competitors and related topics your customers are buzzing about that you may have not considered before
- Value Perception – use Social Listening Tools like Google Analytics to gauge the overall customer opinion of your brand's product or service and whether it can meet their needs
- Setting Social Media Goals
  - which channels and content are performing well, so you can create actionable, realistic social media goals and objectives
- Proving ROI
  - Each time you run a new campaign, monitor your social analytics to see how the content is performing, if people are clicking over to your website and if you're generating new sales.

Reference: Sprout Social

Types of Social Media Analytics

- Performance Analysis
  - Impressions
  - Reach
  - Likes
  - Comments
  - Shares
  - Views
  - Clicks
  - Sales
- Audience Analytics
  - Age
  - Gender
  - Location
  - Device
- Competitor Analysis
  - # of followers
  - Engagement Rate
- Ad Analytics
  - Total number of active ads
  - Clicks
  - Click-through rate
  - Cost-per-click
  - Cost-per-engagement
  - Cost-per-action
  - Conversion rate
  - Total ad spend
- Influencer Analysis
  - Number of posts created per influencer
  - Total number of interactions per post
  - Audience size of each influencer
  - Hashtag usage and engagement
- Sentiment Analysis
  - Track Brand Sentiment
  - Relevant keywords and topics

Solutions from Analytics Companies: Keyhole

- Enterprise-Grade Social Listening & Analytics – for monitoring brand mentions & campaigns
- Publishing & Scheduling – For cross-platform posting
- Profile analytics – for measuring accounts' growth and engagement
- Social Media Trends – for analyzing social media trends
- Influencer Tracking – for measuring influencer campaign performance
- Historical Data – for past analytics, posts and campaigns

Reference: Keyhole

Products from Hootsuite: Social Media Analytics companies

Products

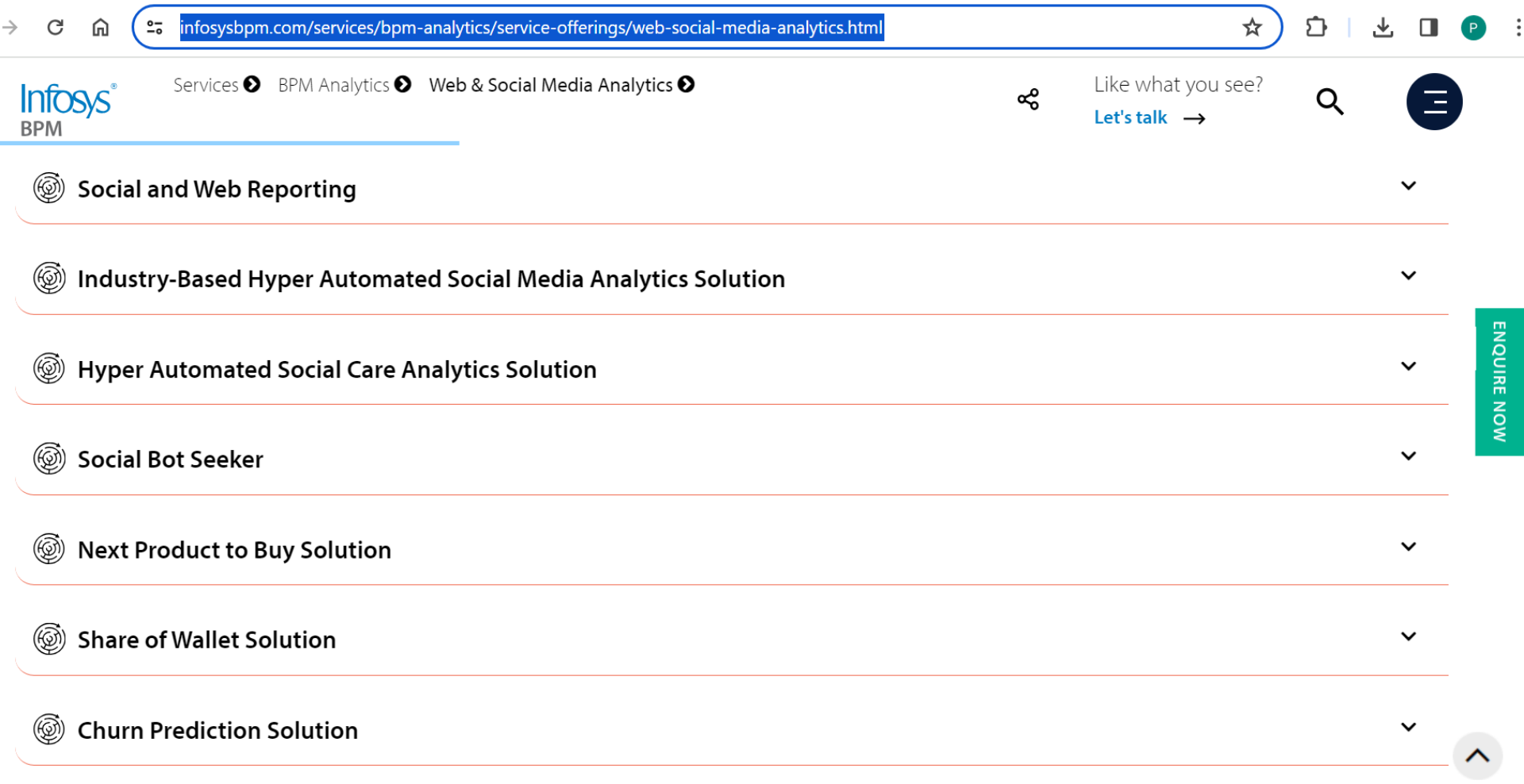
- Publish and schedule
- Engage customers
- Monitor activity
- Advertise content
- Analyze results
- Integrations

Solutions

- Customer Care
- Social Selling
- Employee advocacy
- Social Media Marketing

Reference: https://www.hootsuite.com/

Typical Service Offerings



Source: Infosys Web Site

Unleashing the potential of social media analytics for influencer marketing

Building your Social Media Marketing Strategy for 2022 and beyond

Harnessing the power of social media for marketplace management

Data privacy and ethical considerations in web and social media analytics

The role of web analytics in e-commerce: insights for online retailers

Reasons to use social media analytics

Benefits of social media analytics that you cannot deny

Incorporating web analytics into your marketing strategy

Utilising customer data analysis to boost sales

What is your share of wallet in the market?

Healthcare Data Analytics: Benefits And Use Cases

Social media analytics in 2022

A quick guide to social share of voice

Four ways enterprise analytics will evolve in 2022

Tracking and measuring web and social media analytics

Applying intelligence to analytics: The future of AI in social media

Five emerging trends shaping digital analytics

Digital analytics and its impact on digital marketing

Key benefits of digital analytics for business success

Importance of Social Network Analysis

innovate achieve lead

- **Strategic Advantage**: By understanding the connections and interactions, organisations can mobilise resources more efficiently, enhance cooperation and knowledge sharing, stimulate innovation, and gain a strategic advantage.
- **Improved Understanding**: It paves the way to understanding patterns and trends, uncovering hidden channels of information flow and decision making within and across organisations.
- **Risk Management**: SNA provides a better understanding of dependencies that could pose risks to the functionality and productivity of the system, thereby enhancing risk management.

The fantastic thing about SNA is that it **reveals the invisible** - the behind-the-scenes information flow, the influencers, gatekeepers, and liaisons. By understanding this, **businesses can enhance their strategies, communications and understand the informal and formal structures within their organisation.**

Reference: Study Smarter

Social Network Analysis Methods

- **Centrality Measures**: This gives insights into the most influential or central actors in a network.
- **Clique Analysis**: This helps identify sub-groups of nodes that are more densely connected to each other than to other nodes in the network.
- **Ego Network Analysis**: This focuses on a single node (the ego) and the nodes to which it is directly connected (the alters).
- **Cohesion Measures**: These measure how tightly knit a network is, helping you understand the strength or weakness of the overall network cohesion.

Reference: Study Smarter

Centrality Measures

- A person with a large number of friends on a social media platform would have high Degree Centrality. However, having a lot of friends doesn't necessarily mean that a person can reach others quickly, as their friends may not be well connected
- Closeness Centrality is a measure of how fast information can flow from a given node to other nodes in the network. Mathematically speaking, it is the reciprocal of the sum of the shortest paths from a node to all other nodes.
- Betweenness Centrality is a measure of the extent to which a node lies on paths between other nodes. Nodes with high betweenness centrality serve as a bridge (or a 'broker') from one part of a network to another.

Reference: Study Smarter

SNA Examples

- Understanding political structures
- Investigating the spread of diseases
- Tracing the flow of information in an organisation
- Transaction web in cryptocurrencies

Within a corporate setting:

- Insights into the informal networks that exist alongside the official organisation chart. For instance, employees often seek guidance not from their official superiors but from experienced colleagues.
- An SNA in this scenario could help to identify these individuals, measure their importance (using measures like degree centrality and betweenness centrality), and assess the impact of their eventual retirement or departure from the company.
- SNA could also showcase structural gaps where communication or collaboration is missing but necessary.

Reference: Study Smarter

SNA in Marketing

In marketing and brand strategy, SNA can help chart the landscape of social influencers. By determining the degree centrality, one can identify individuals who, due to their vast network of connections, can be instrumental in spreading content widely.

Betweenness centrality, on the other hand, can help identify those individuals who serve as critical brokers or bridges between diverse parts of the network. They might not have the highest number of connections, but they hold influence because they link different communities or groups.

A cosmetics company planning to release a new product might use SNA to identify key influencers in the beauty community. By sending products to these individuals and securing their endorsement, the company can ensure that news of the product reaches a wide audience more effectively than through traditional advertising methods.

Reference: Study Smarter

Other SNA Applications

- **Sociology**: Just as the name suggests, SNA was first developed by sociologists to understand social structures. It can unveil the complexities of human interactions, such as analysing online communities, tracking socioeconomic disparity, and studying the diffusion of cultural trends.
- **Computer Science & IT**: SNA has become a vital part of computational data analysis, primarily for the Internet and its structure. It's employed in areas like web graph analysis, cybersecurity for tracing the proliferation of malware and even optimising cloud computing networks.
- **Political Studies**: In political science, SNA is used to study policy networks, political parties, political blogs, or even to understand power structures among nations. It also aids in tracking the diffusion of political ideologies and trends.
- **Business Operations**: SNA is actively utilised to optimise organisational structures, enhance communication networks, and improve marketing strategies

Reference: Study Smarter

SNA in Business Operations

Social Network Analysis emerges as a powerful process enhancement tool. It offers a unique perspective to aid solving many business-related issues, like enhancing team collaborations, improving inter-departmental communication or even understanding customer behaviours.

Employee Interaction and Collaboration

Organisations are essentially a complex web of interactions and relationships. SNA helps to visualise this web, further enabling the organisation to understand the communication flow and thereby, promoting better collaborations. Using measures like degree centrality and betweenness centrality, one can identify key individuals who are acting as information gatekeepers.

Example: Suppose there's an individual who doesn't have an official leadership title, but their departure greatly hampers the workflow. This could possibly be because they hold a pivotal position within the informal network, answering colleagues' queries, mediating discussions, or ensuring coordination. Understanding these informal roles through SNA could significantly enhance workflow management.

Reference: Study Smarter

SNA in Business Operations

Organisational Knowledge Management

Knowledge and information in an organisation do not follow a clear-cut path as depicted by official hierarchies. Instead, it flows across organisational boundaries in rather unexpected ways. SNA allows the identification of such unconventional paths.

Example

‘T-shaped' skills, for instance, where a person has depth of knowledge in one subject (the vertical bar of the T) along with the ability to collaborate across disciplines and apply knowledge in areas of expertise other than their own (the horizontal bar of the T), are essential for innovation. SNA can help identify such individuals with 'T-shaped' skills and foster cross-disciplinary learning.

Reference: Study Smarter

SNA in Business Operations

Consumer Behaviour Analysis

On the marketing front, SNA can help understand consumer behaviours, preferences, and their decision-making process. By studying consumer networks, organisations can identify influences that impact purchasing decisions or track the diffusion of new product knowledge. With this, companies can serve more targeted advertisements and understand the potential buyer's journey.

Reference: Study Smarter

Advantages of SNA

Uncovering Hidden Relationships: The complexity of the relational data analysed is often such that making sense out of it is rather challenging. Social Network Analysis with its computational methods allows you to unravel hidden relationships and dynamics within a network, something not easily attainable through traditional data investigation techniques.

Enhanced Predictability: By determining the centrality measures (like degree, closeness, and betweenness centrality), you can predict emerging trends and behaviors within a network. In a business scenario, such predictability could enable better marketing strategies and target-specific operations.

Visualisation: One of the major advantages of SNA is its capacity to visually present complex data in an understandable form. This visual representation aids in the easy recognition of patterns, key players, and relationships.

Robustness in Various Fields: As discussed in previous sections, SNA is adept at managing multifaceted network problems in fields as diverse as business operations, sociology, computer science, and politics, among others, allowing it to adapt to a wide range of data contexts.

Reference: Study Smarter

Disadvantages of SNA

Data Entry Difficulties: The process of converting network data for SNA can be challenging and time-consuming. Collecting relational data can also prove to be more demanding than gathering simple attribute data as you need to account for connections and not just properties.

Data Privacy Concerns: With the rise of data privacy awareness, issues concerning the privacy of the network's members can pose significant challenges. Consent, purpose limitation, and data minimisation are all significant hurdles when analysing network data, especially those of a more personal nature (social networks, for example).

Interpretation Challenges: While visualisation helps represent data, SNA's interpretation is still complex due to inherent network complexity. Mistaking correlation for causality is a common issue in Social Network Analysis.

Dynamic Nature: Networks are constantly evolving and changing over time. Capturing a snapshot of the network at a single time point may therefore not provide an accurate representation as the state of the network could shift rapidly.

Reference: Study Smarter

SNA in Supply Chain Management

innovate achieve lead

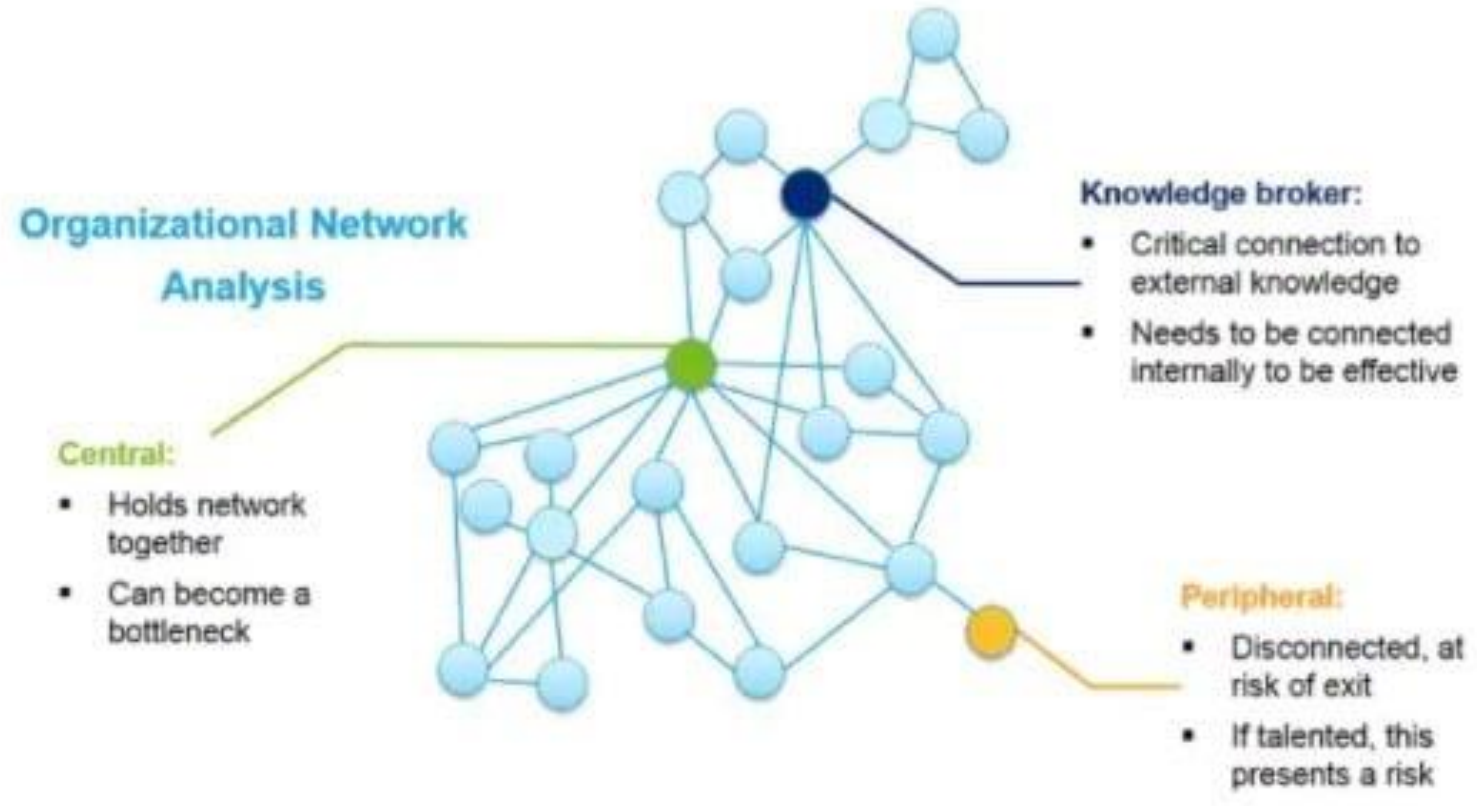
- A supply chain can be modeled into a network of supplier/consumer relations. Network analysis on the supply chain helps us improve the operation efficiency by identifying and eliminating less important nodes (suppliers/warehouses). It can help identify crucial nodes in the network and create a standby in crises or emergencies.

- Nodes include Retailers, Suppliers, Warehouses, Transporters, Regulatory agencies.

- SNA applications can help manufacturers identify more operationally critical nodes and identify potential sources to increase the number of connections to suppliers. This can also help identify any bottlenecks in the supply process and inventory management.

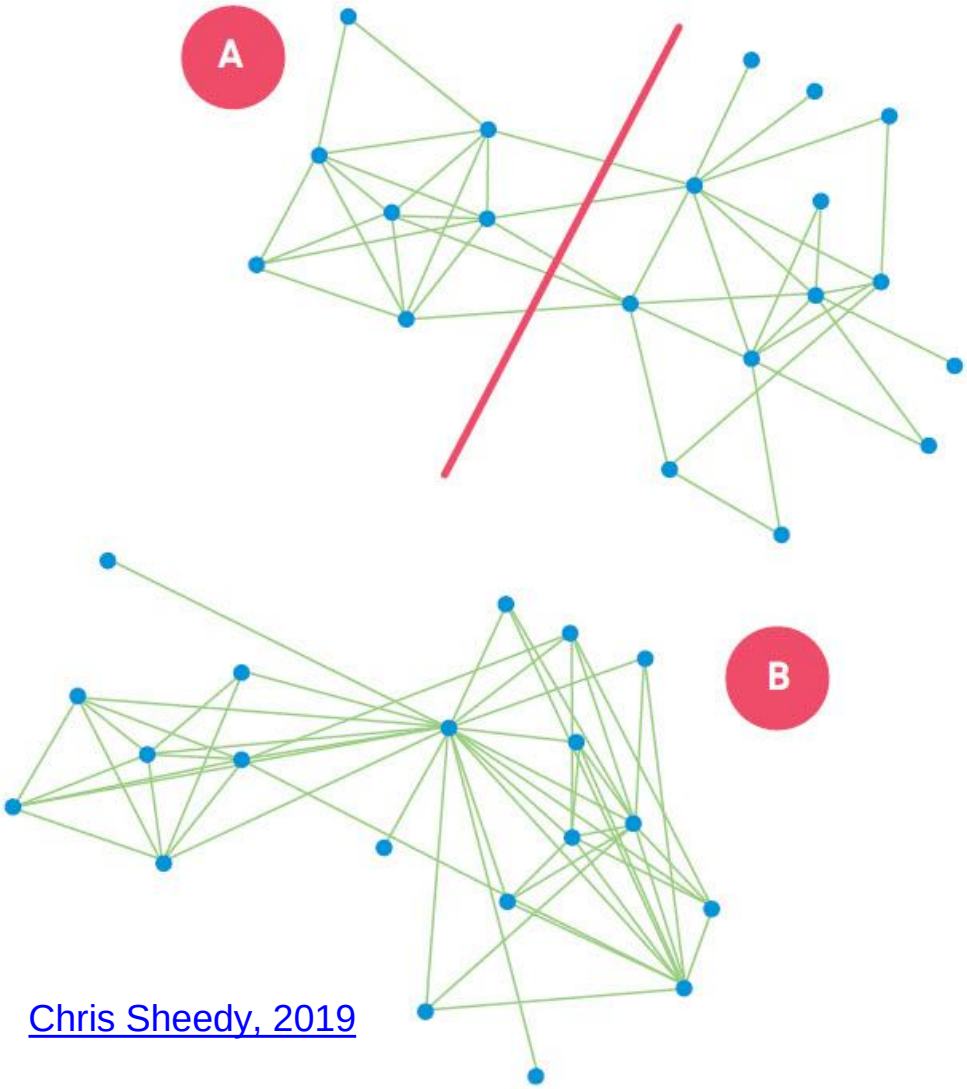
Source: https://www.latentview.com/blog/a-guide-to-social-network-analysis-and-its-use-cases/

Organizational Network Analysis



Source: https://www2.deloitte.com/us/en/pages/human-capital/articles/organizational-network-analysis.html

SNA in Teams



- Scenario (A), the group was split into two subgroups that were relatively comfortable communicating with each other.
- The group had, in fact, been two teams, brought together under a single manager. They were co-located, but still largely working as two separate groups.
- The company's management thought team building might help bring the subgroups together.
- Scenario (B) shows the team three months later. There were more connections within the group, with one individual particularly pivotal in unifying the team.
- What caused this change? Not traditional team-building exercises, but "targeted self-disclosure exercises".

Chris Sheedy, 2019

Building Trust Through Skillful Self-Disclosure

SNA in Medicine

Day nurse 1 > Night nurse 1	never gave info (0)
Day nurse 1 > Night nurse 2	often gave info (3)
Day nurse 1 > Night nurse 3	constantly gave info (4)
etc. for all night nurses	
Day nurse 2 > Night nurse 1	seldom gave info (1)
Day nurse 2 > Night nurse 2	seldom gave info (1)
Day nurse 2 > Night nurse 3	constantly gave info (4)
etc. for all night nurses	

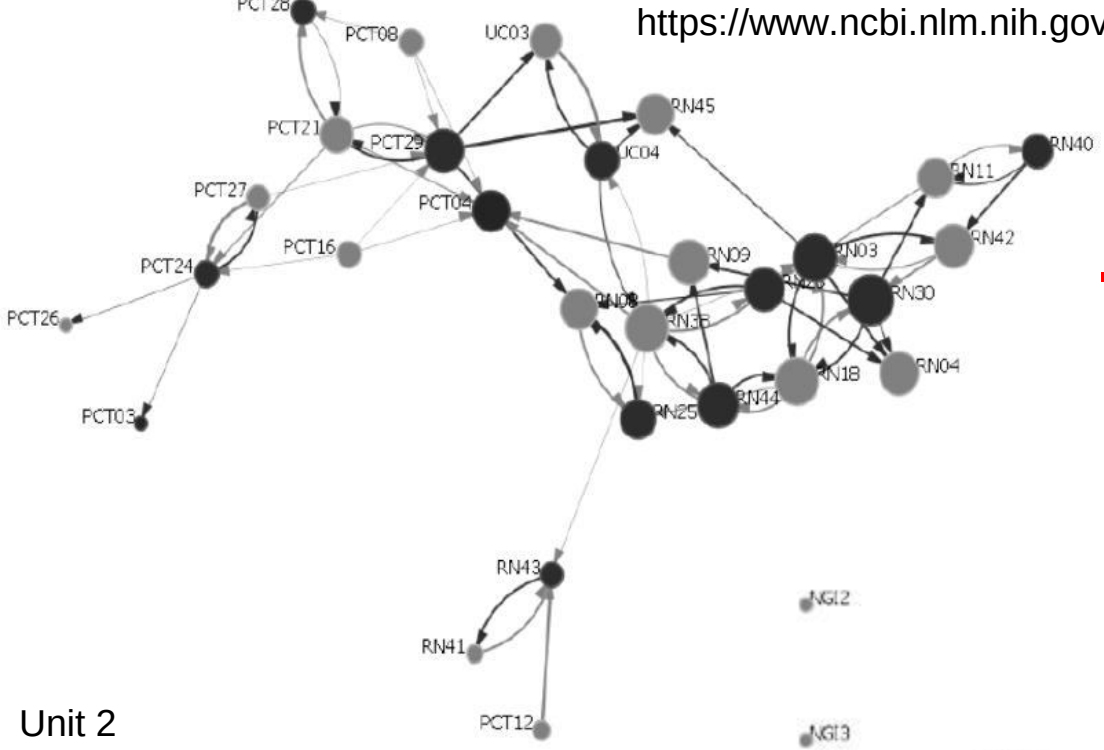
then etc. for all day nurses

https://www.statistics.com/social-network-analysis-sna-in-medicine/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4156878/

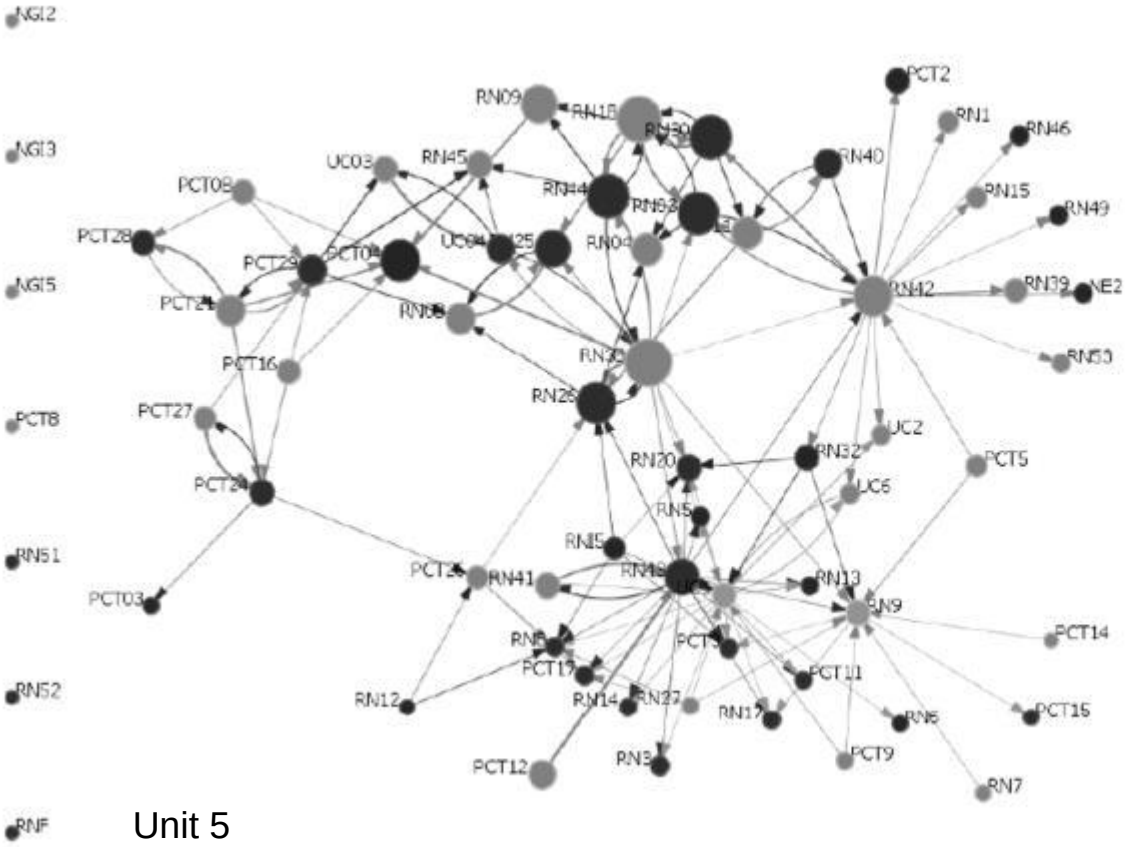
innovate achieve lead

- PCT: Patient Care Technician
- RN: Registered Nurse
- Gray circles: Day shift
- Dark circles: Night shift
- Thicker links: Higher comm frequency
- Larger node size represents higher Eigenvector Centrality values



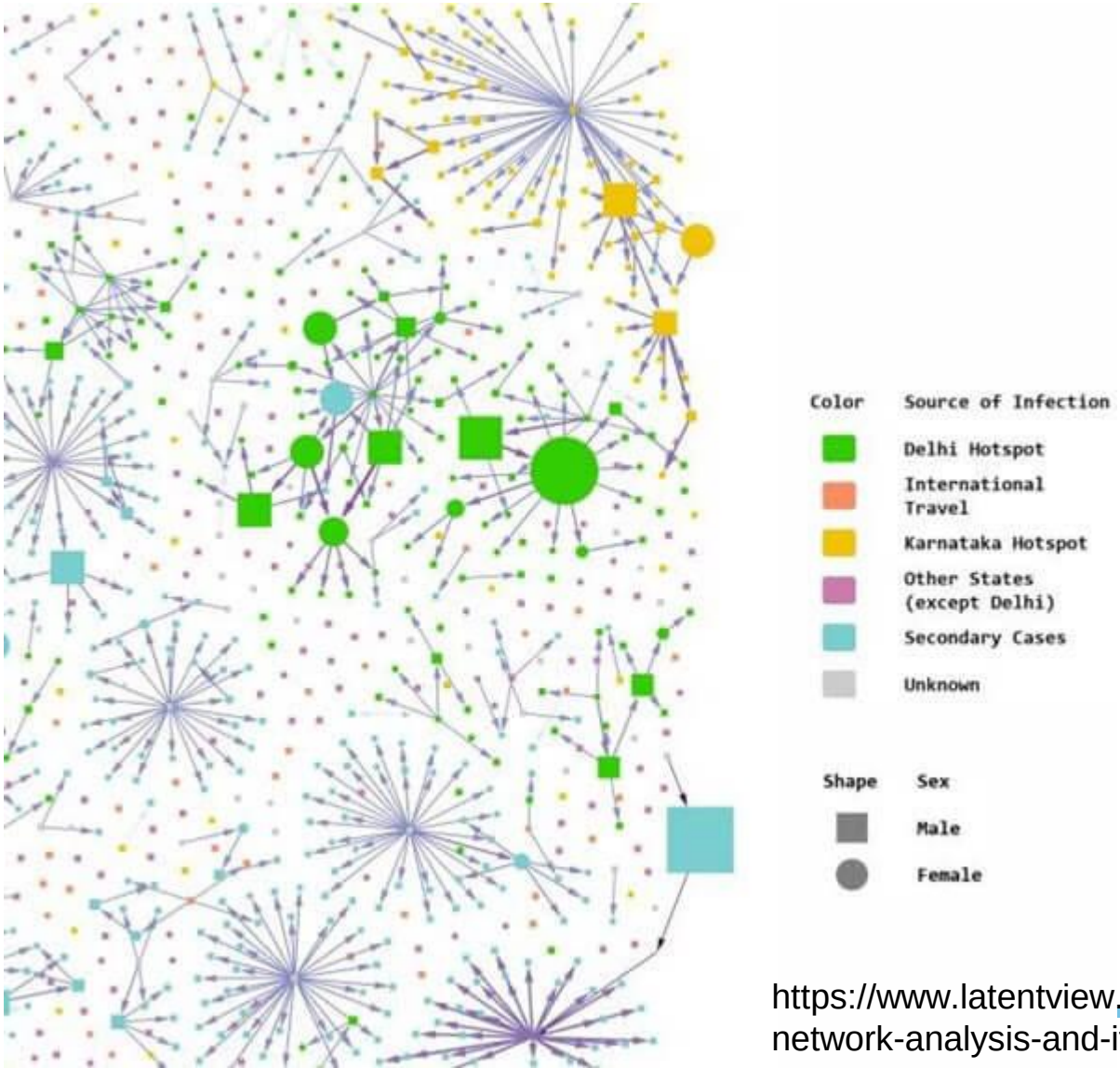
Unit 2

- Unit 2's diffusion metric is nearly twice as high (.43) as that for the much larger Unit 5 (.28). Information might diffuse quickly across the network of Unit 2
- The average Eigenvector Centrality value for Unit 2 is .45, compared with that of Unit 5 (.20). This suggests that on the smaller unit, more individuals are connected to highly connected staff.
- In Unit 5, two of the more influential RNs (high Eigenvector Centrality) are not communicating with staff not on their shift, and there are a number of "pendants" (people with single links). The pendants are usually PCTs.



Unit 5

SNA for Infectious Disease Tracking



https://www.latentview.com/blog/a-guide-to-social-network-analysis-and-its-use-cases/

Diffusion of microfinance

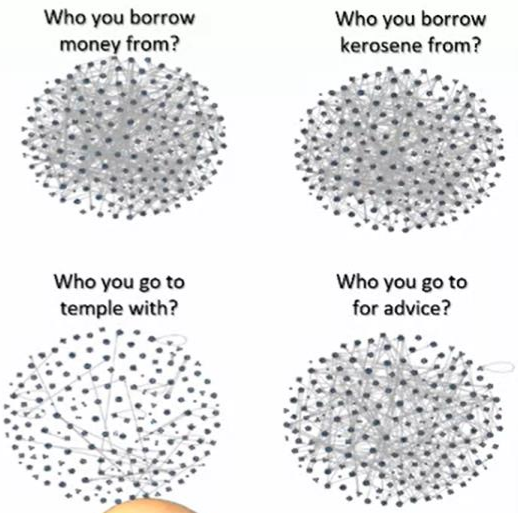
Centrality Application

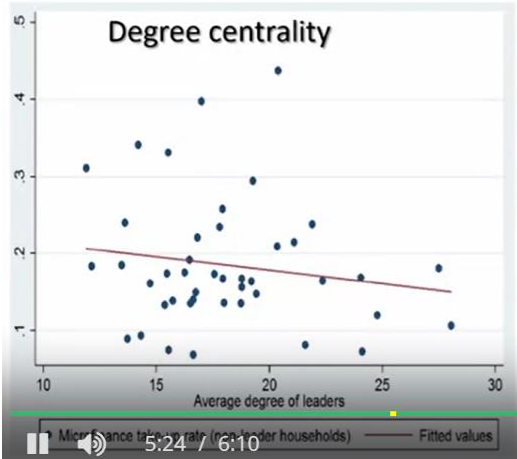
➢ **Diffusion of microfinance**

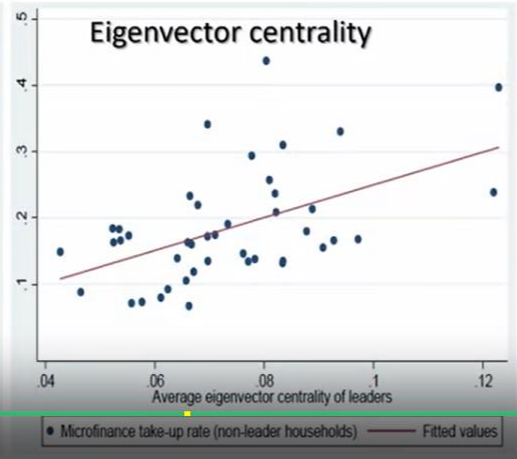
Banerjee, Chandrasekhar, Duflo & Jackson (2013). The Diffusion of Microfinance. *Science*, *341*(6144), 1236498. https://doi.org/10.1126/science.1236498

- 75 rural villages in Karnataka/ India, without microfinance
- Bank entered 43 of them and offered microfinance

- Question:
  Who to contact first to spread the innovation?
- Challenge:
  How to map the network: "who would you borrow from?"
  …they created 13 different/ multiplex networks…



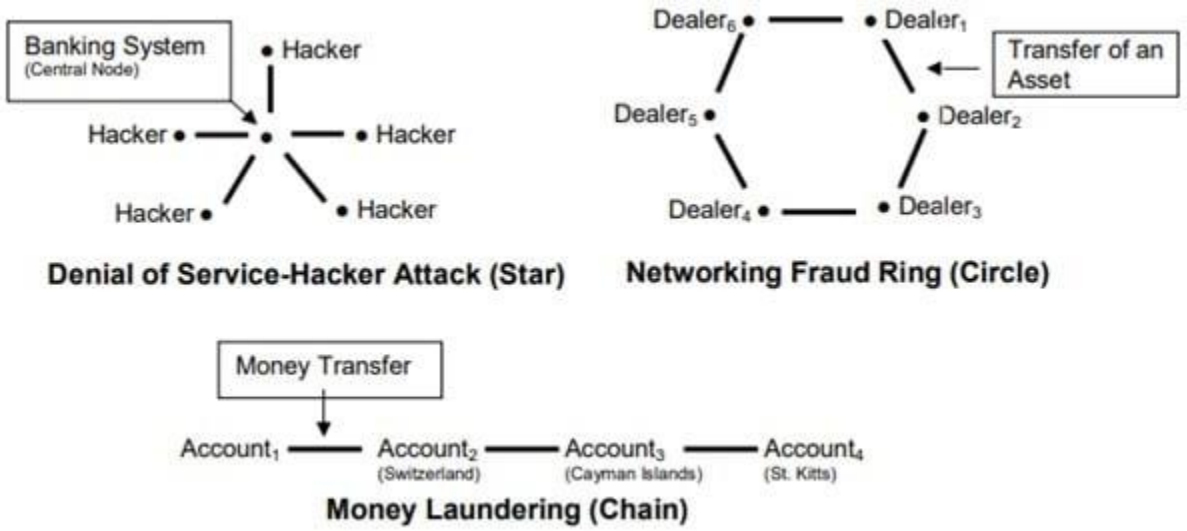




=> contact those whose friends have many friends!

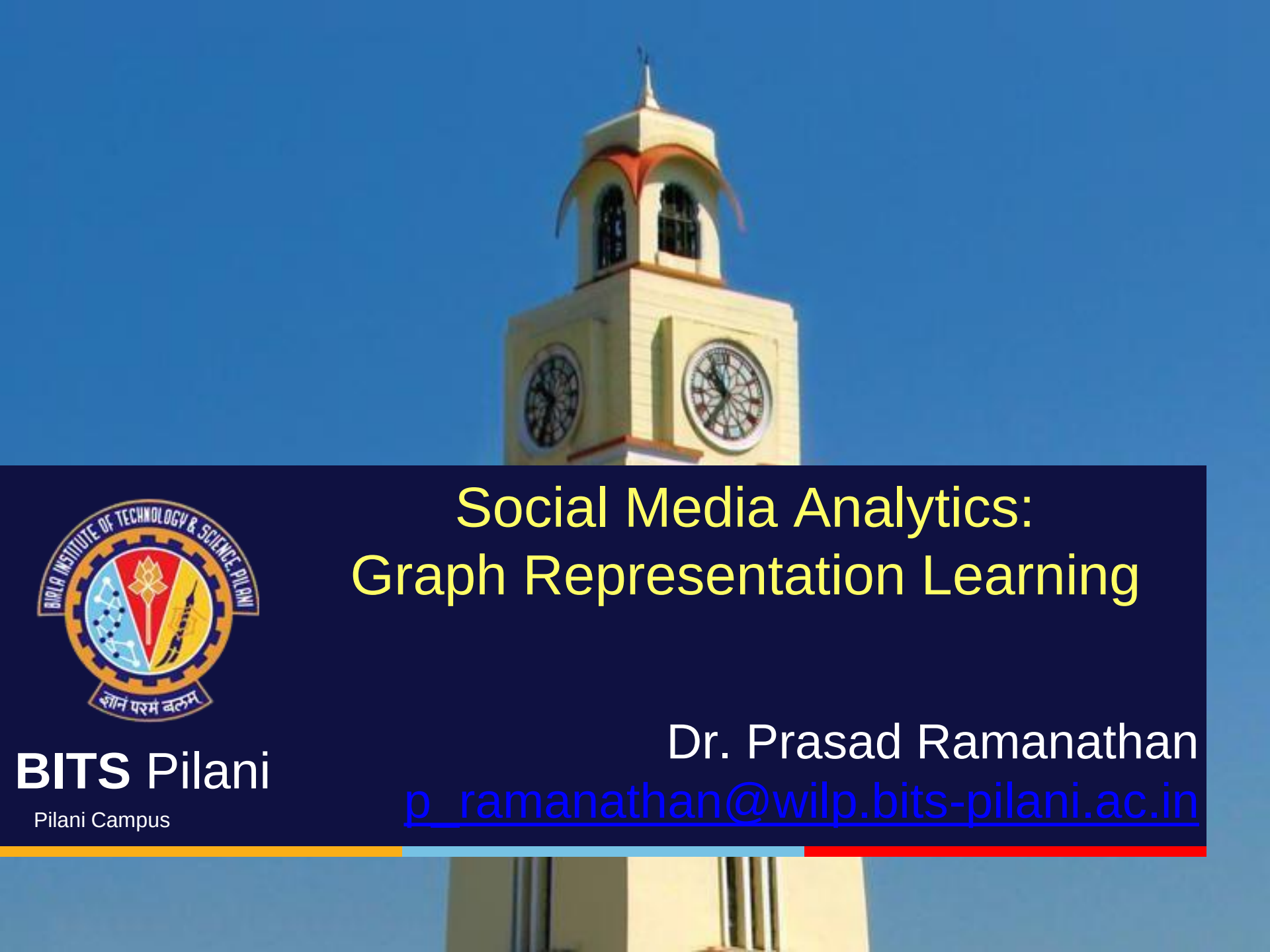
Reference: https://www.coursera.org/learn/social-network-analysis/home/week/2

SNA in Finance / Fraud Detection



https://www.latentview.com/blog/a-guide-to-social-network-analysis-and-its-use-cases/

Thank you



Social Media Analytics: Graph Representation Learning

Dr. Prasad Ramanathan

p_ramanathan@wilp.bits-pilani.ac.in

BITS Pilani

Pilani Campus

References

Acknowledgments:

1. Content provided by Prof. Jure Leskovec and others teaching the following course is gratefully acknowledged
   - CS224W: Machine Learning with Graphs
   - Stanford University
   - http://cs224w.Stanford.edu

2. Graph Representation Learning by William L. Hamilton

3. Tanmay Chakraborty, Social Network Analysis, Chapter 9, https://www.youtube.com/playlist?list=PLqGkIjcOyrGnu-v7IKyDd64Y-gd0gNlqK

Why Graphs?

Graphs are a general language for describing and analyzing entities with relations/interactions

Graphs: Machine Learning

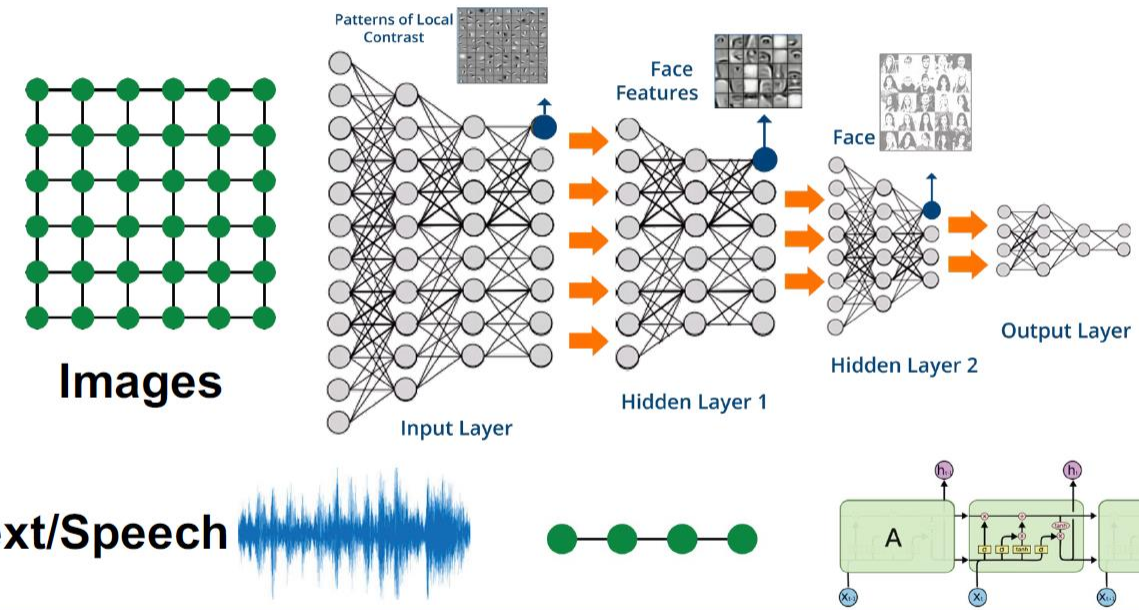
Complex domains have a rich relational structure, which can be represented as a **relational graph**

By explicitly modeling relationships we achieve better performance!

Main question:

How do we take advantage of relational structure for better prediction?

Today: Modern ML Toolbox



Modern deep learning toolbox is designed for simple sequences & grids

Doubt thou the stars are fire,
Doubt that the sun doth move,
Doubt truth to be a liar,
But never doubt I love...

Text

Audio signals

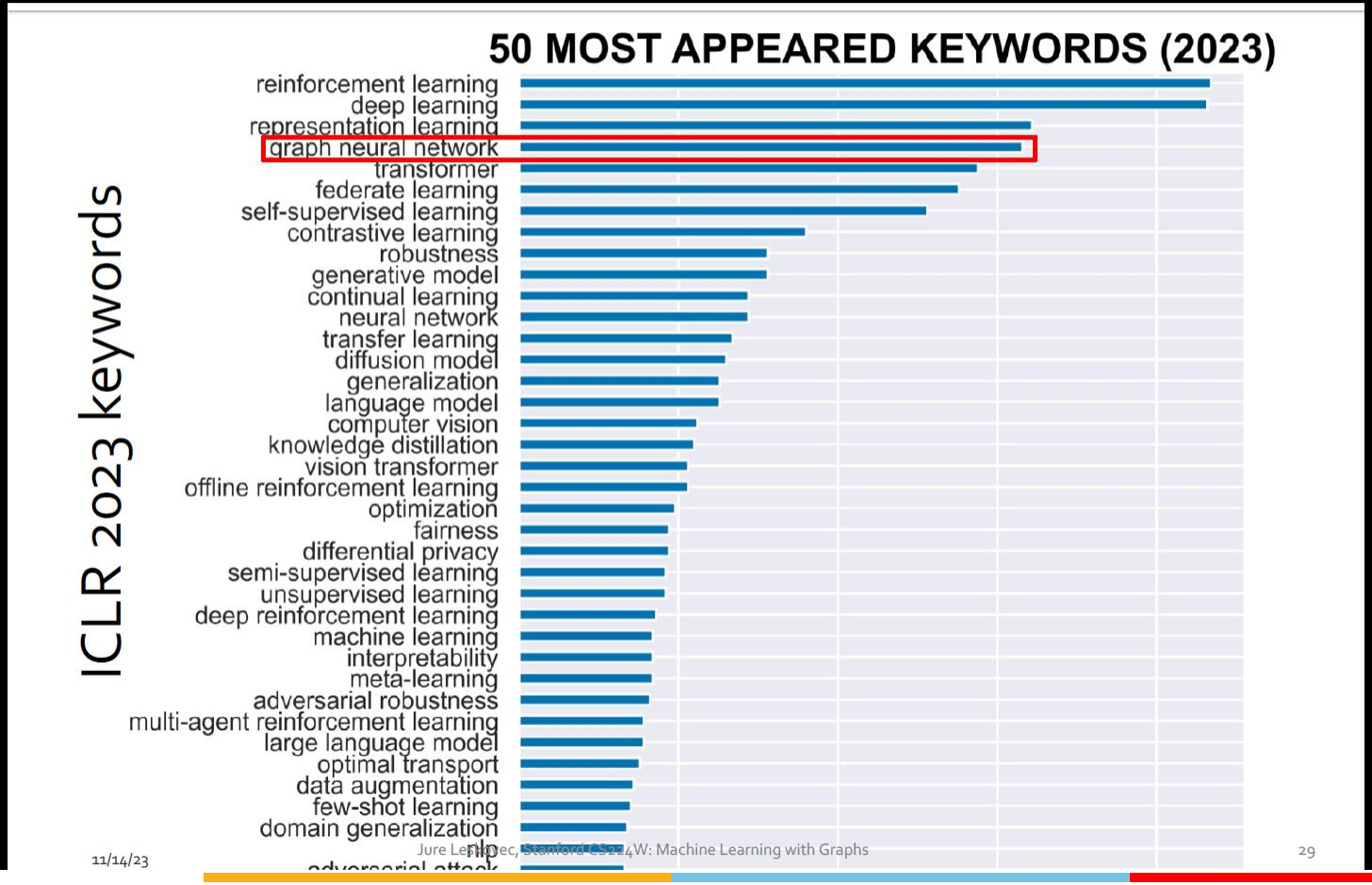
Images

Modern
deep learning toolbox
is designed for
sequences & grids

How can we develop neural networks that are much more broadly applicable?

Graphs are the new frontier of deep learning

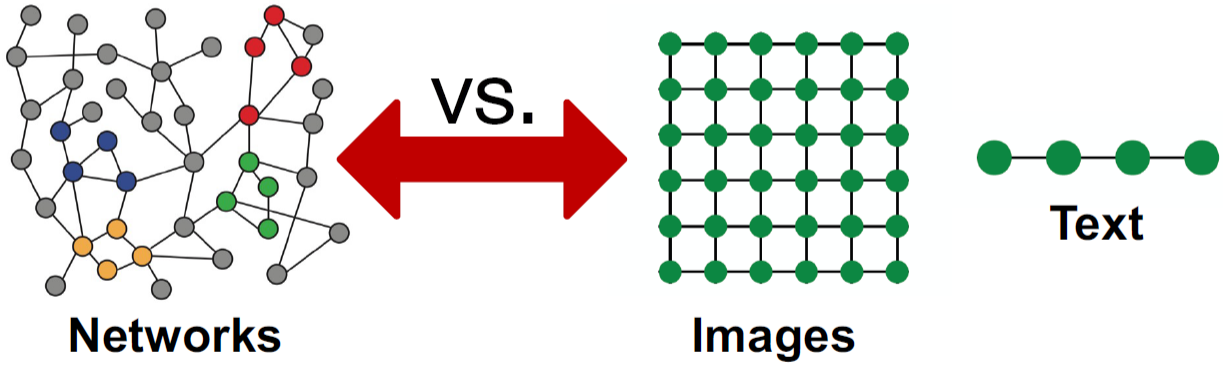
Hot subfield in ML



Why is Graph Deep Learning Hard?

Networks are complex.

- Arbitrary size and complex topological structure (*i.e.*, no spatial locality like grids)



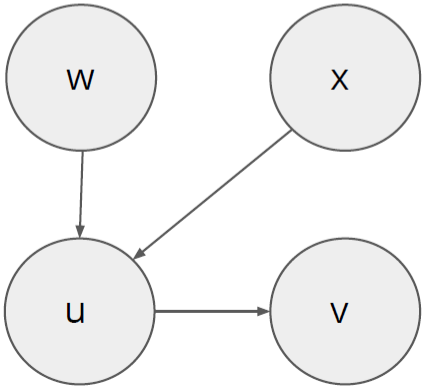
- No fixed node ordering or reference point
- Often dynamic and have multimodal features

Graph encoding as a matrix

Adjacency Matrix: $\mathbf{A} \in \mathbb{R}^{|V| \times |V|}$

- In this example, binary matrix encoding of a unweighted graph
- Rows/columns number the nodes, matrix elements encode edges

$V = \{u, v, w, x\}$; $E = \{(w, u), (x, u), (u, v)\}$



$\mathbf{A} =$

(from) \ (to)	u	v	w	x
u	0	**1**	0	0
v	0	0	0	0
w	**1**	0	0	0
x	**1**	0	0	0

Do the matrices encode the same graph?

0	**1**	0	0
0	0	0	0
1	0	0	0
1	0	0	0

0	0	0	0
0	0	**1**	0
1	0	0	0
0	0	**1**	0

Hint: Have we given you enough information?

They are the same encoding!

	u	v	w	x
u	0	**1**	0	0
v	0	0	0	0
w	**1**	0	0	0
x	**1**	0	0	0

	v	w	u	x
v	0	0	0	0
w	0	0	**1**	0
u	**1**	0	0	0
x	0	0	**1**	0

Considerations for GNN

1) Nodes are not i.i.d* (we are modeling an interconnected set of nodes)

2) A NN modeling a graph should be permutation invariant and equivariant

   - Adjacency matrix orders nodes arbitrarily

For permutation matrix **P**, function f that takes in an adjacent matrix **A**:

Permutation Invariance Property: $f(\mathbf{P}\mathbf{A}\mathbf{P}^{\mathbf{T}}) = f(\mathbf{A})$

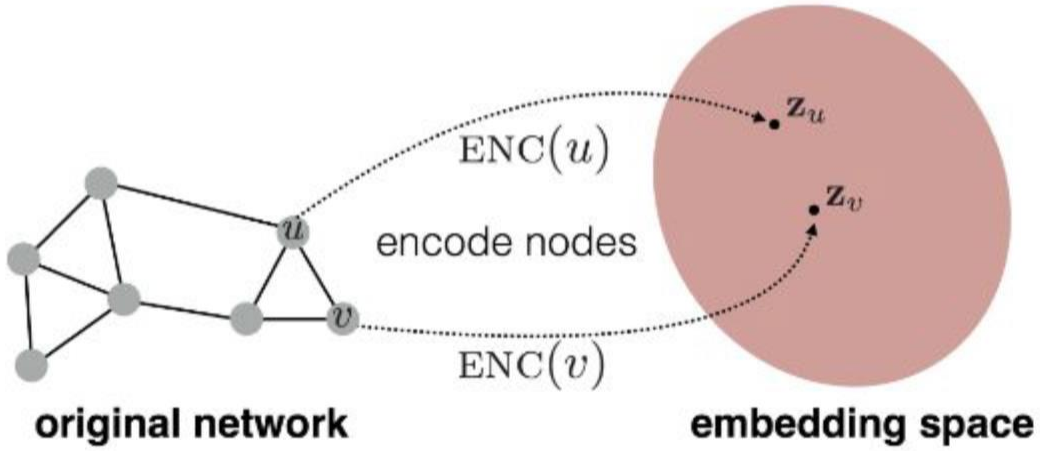
Permutation Equivariance Property: $f(\mathbf{P}\mathbf{A}\mathbf{P}^{\mathbf{T}}) = \mathbf{P}f(\mathbf{A})$

*i.i.d = independent and identically distributed

Ref: Everett Knag, Justin Saluja, Chaitanya Srinivasan, Prakarsh Yadav, “Graph Neural Networks”, 11-785 Deep Learning Spring 2021

Considerations for GNN

3) Find an encoding that preserves the graph structure

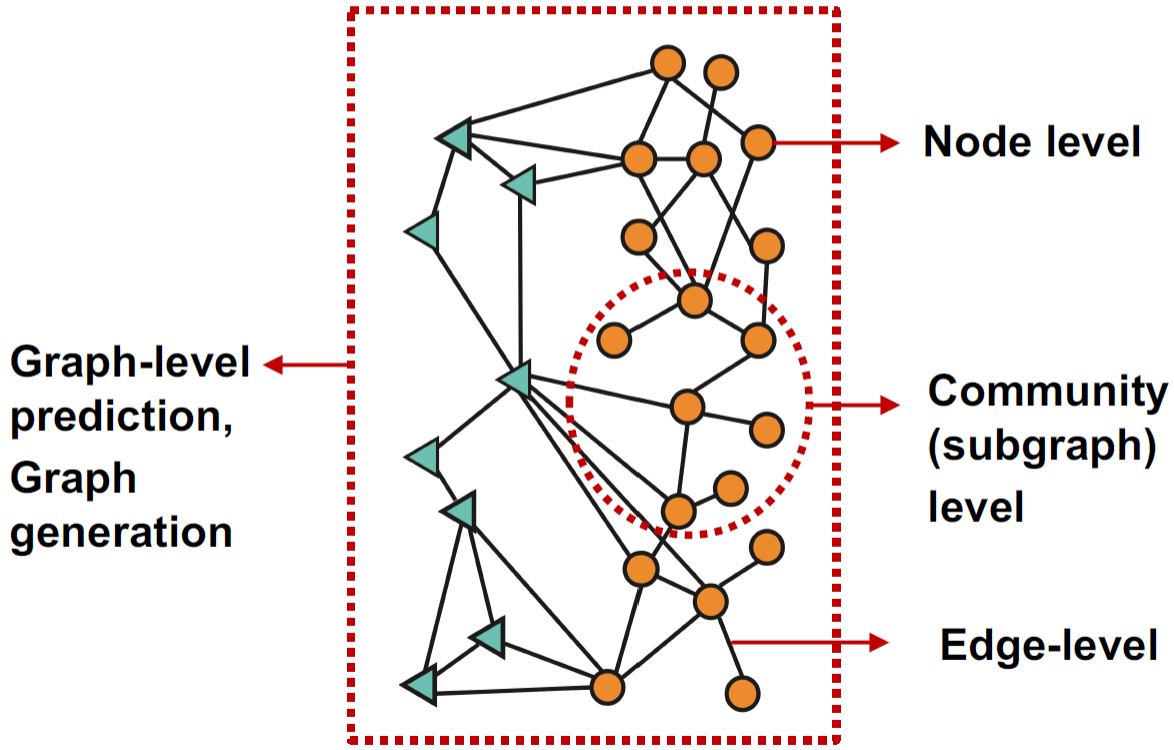


Insight: exploit homophily - a neighborhood of nodes tend to have shared attributes

Figure 3.1, *Graph Representation Learning*

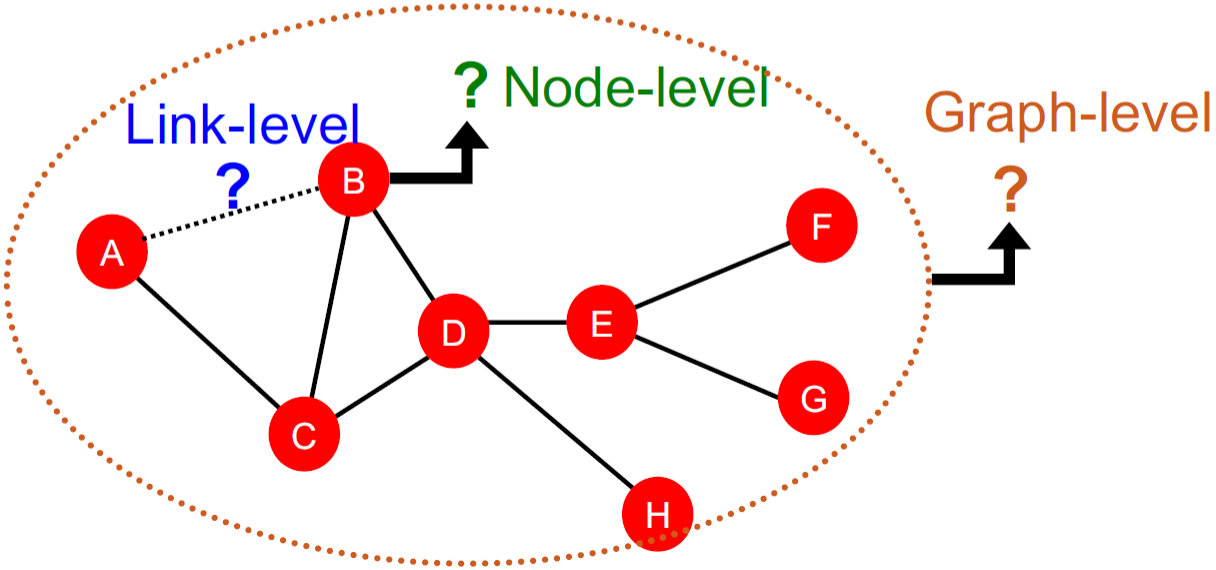
Applications of Graph ML

Different Types of Tasks



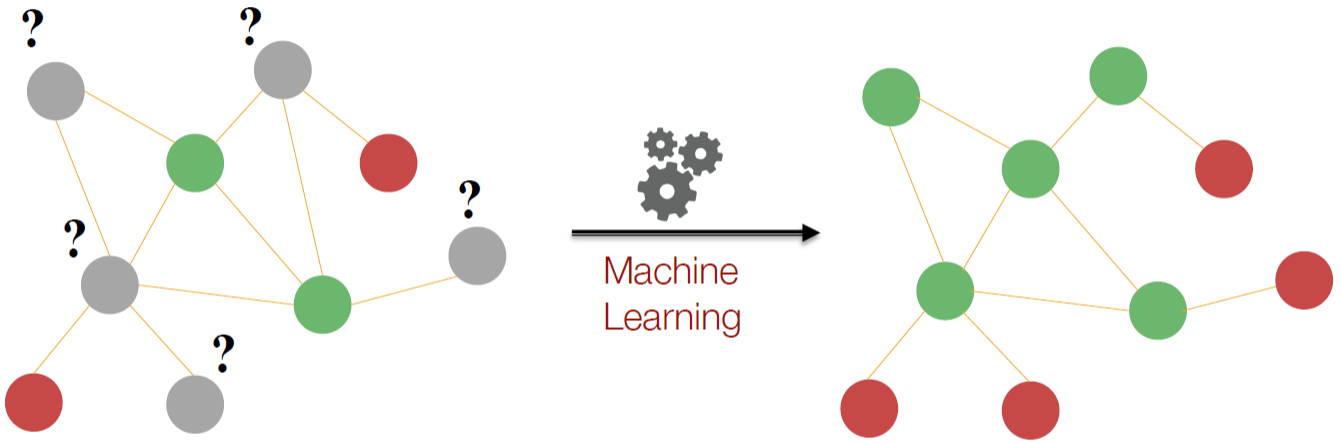
Machine Learning Tasks: Review

- Node-level prediction
- Link-level prediction
- Graph-level prediction



Node-level Predictions

Node-Level Tasks

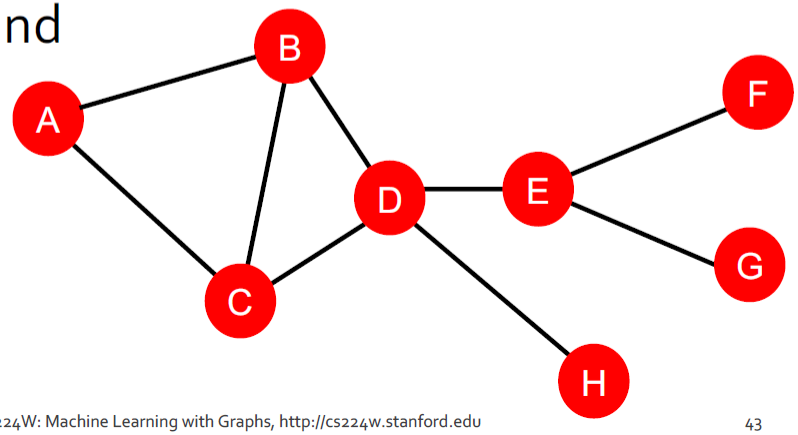


Node classification

Node-Level Network Structure

Goal: Characterize the structure and position of a node in the network:

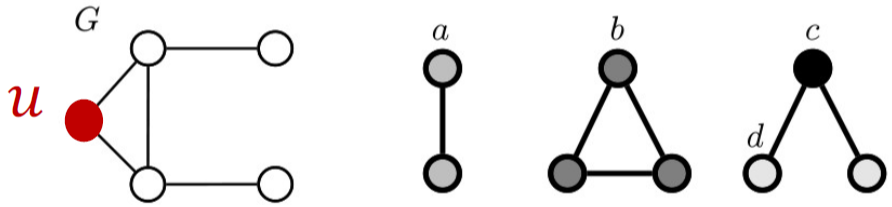
- Node degree
- Node importance & position
  - E.g., Number of shortest paths passing through a node
  - E.g., Avg. shortest path length to other nodes
- Substructures around the node



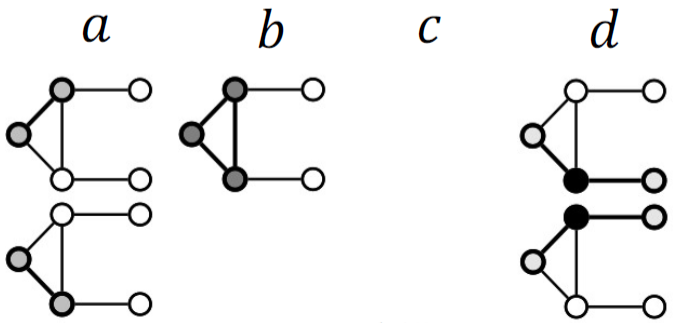
Node's Subgraphs: Graphlets

- **Graphlets:** A count vector of rooted subgraphs at a given node.
- **Example:**

All possible graphlets on up to 3 nodes



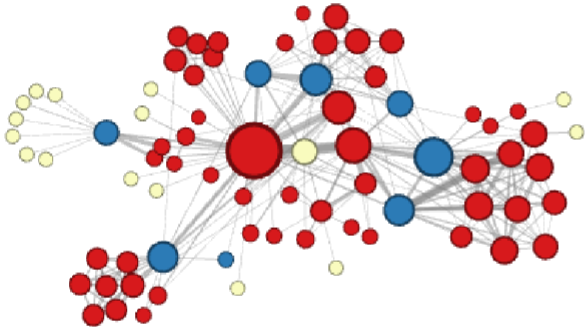
Graphlet instances of node u:



Graphlets of node u:
a, b, c, d
[2,1,0,2]

Discussion

Different ways to label nodes of the network:

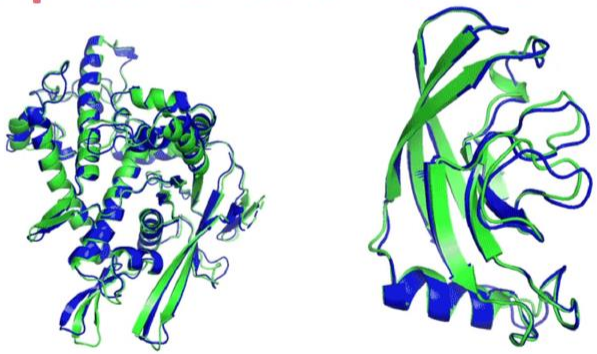


Node features defined so far would allow to distinguish nodes in the above example

However, the features defines so far would not allow for distinguishing the above node labelling

Example (1): Protein Folding

Computationally predict a protein's 3D structure based solely on its amino acid sequence:
For each node predict its 3D coordinates



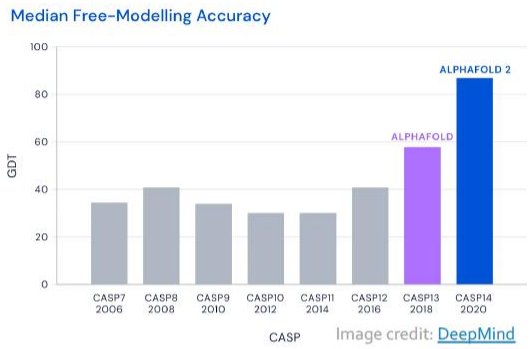
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

Image credit: DeepMind

AlphaFold: Impact



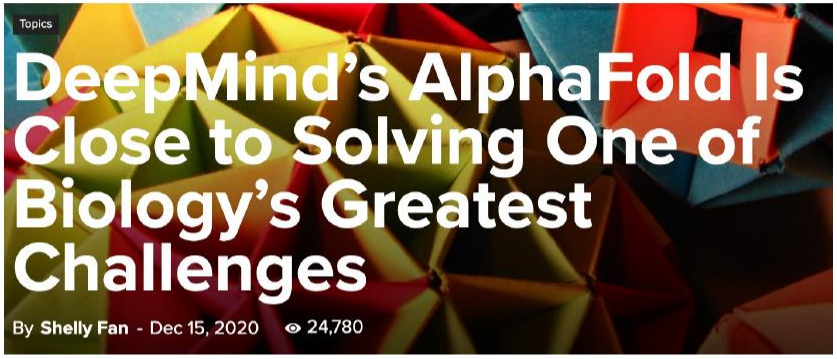


Image credit: SingularityHub

AlphaFold's AI could change the world of biological science as we know it

DeepMind's latest AI breakthrough can accurately predict the way proteins fold

Has Artificial Intelligence 'Solved' Biology's Protein-Folding Problem?

12-14-20

DeepMind's latest AI breakthrough could turbocharge drug discovery

AlphaFold: Solving Protein Folding

- **Key idea:** "Spatial graph"
  - **Nodes:** Amino acids in a protein sequence
  - **Edges:** Proximity between amino acids (residues)

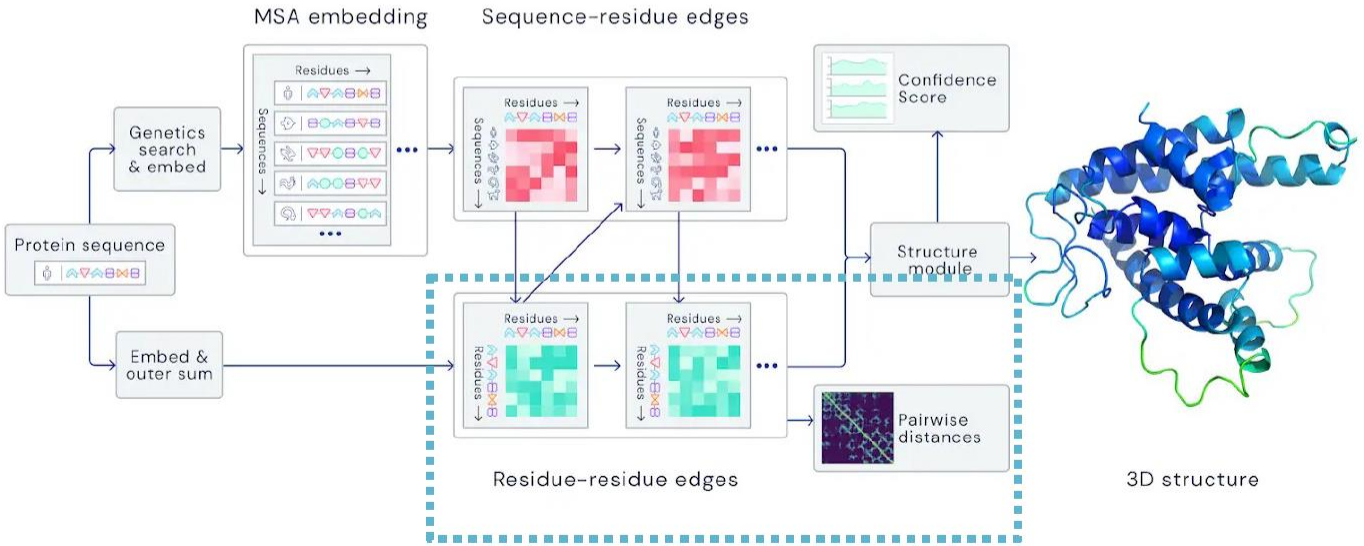
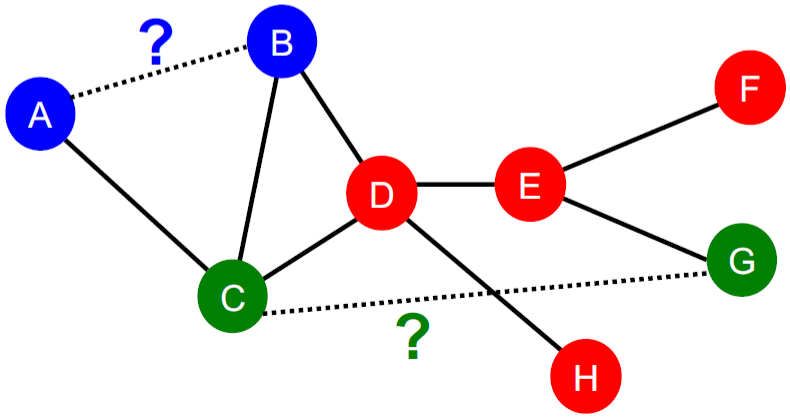


Image credit: DeepMind

Link Prediction

Link-Level Prediction Task

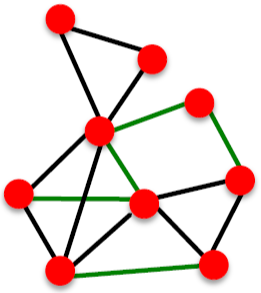
- The task is to predict **new/missing/unknown links** based on the existing links.
- At test time, node pairs (with no existing links) are ranked, and top K node pairs are predicted.
- Task: Make a prediction for **a pair of nodes**.



Link Prediction as a Task

Two formulations of the link prediction task:

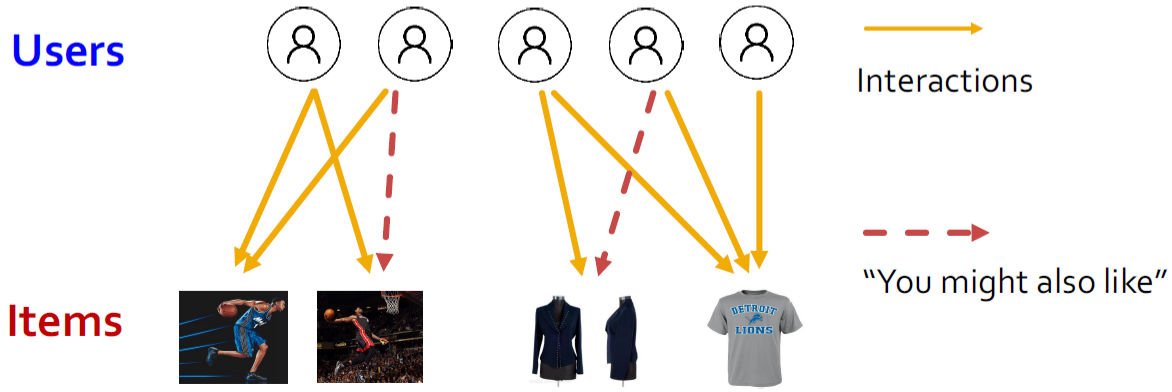
- **1) Links missing at random:**
  - Remove a random set of links and then aim to predict them
- **2) Links over time:**
  - Given $G[t_0, t_0']$ a graph defined by edges up to time t_0', **output a ranked list L** of edges (not in $G[t_0, t_0']$) that are predicted to appear in time $G[t_1, t_1']$
  - **Evaluation:**
    - $n = |E_{new}|$: # new edges that appear during the test period $[t_1, t_1']$
    - Take top n elements of L and count correct edges



$G[t_0, t_0']$
$G[t_1, t_1']$

Example (1): Recommender Systems

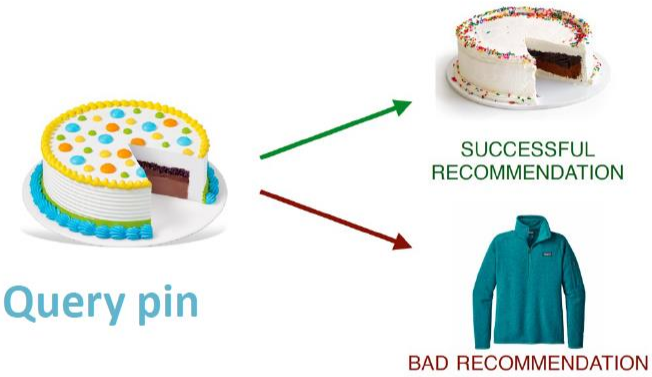
- **Users interacts with items**
  - Watch movies, buy merchandise, listen to music
  - **Nodes:** Users and items
  - **Edges:** User-item interactions
- **Goal: Recommend items users might like**



Ying et al., Graph Convolutional Neural Networks for Web-Scale Recommender Systems, KDD 2018

PinSage: Graph-based Recommender

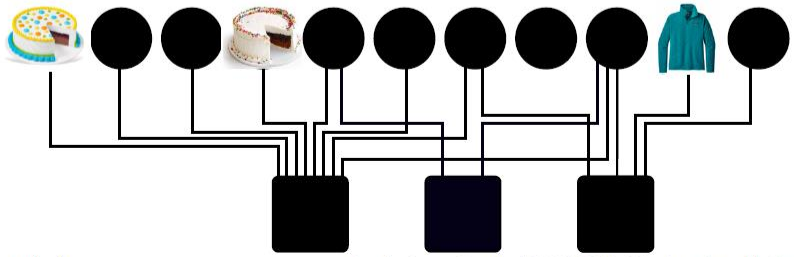
Task: Recommend related pins to users

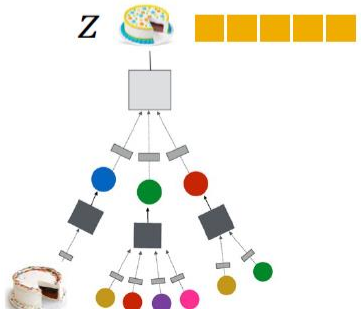


Task: Learn node embeddings z_i such that

$$d(z_{cake1}, z_{cake2}) < d(z_{cake1}, z_{sweater})$$

Predict whether two nodes in a graph are related



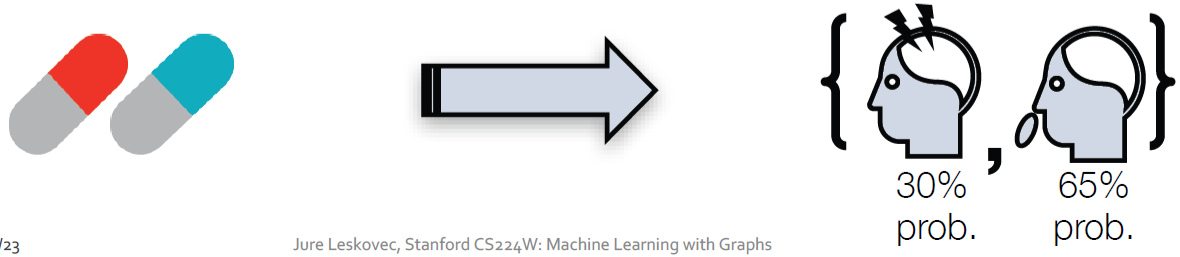


Example (2): Drug Side Effects

Many patients take multiple drugs to treat complex or co-existing diseases:

- 46% of people ages 70-79 take more than 5 drugs
- Many patients take more than 20 drugs to treat heart disease, depression, insomnia, etc.

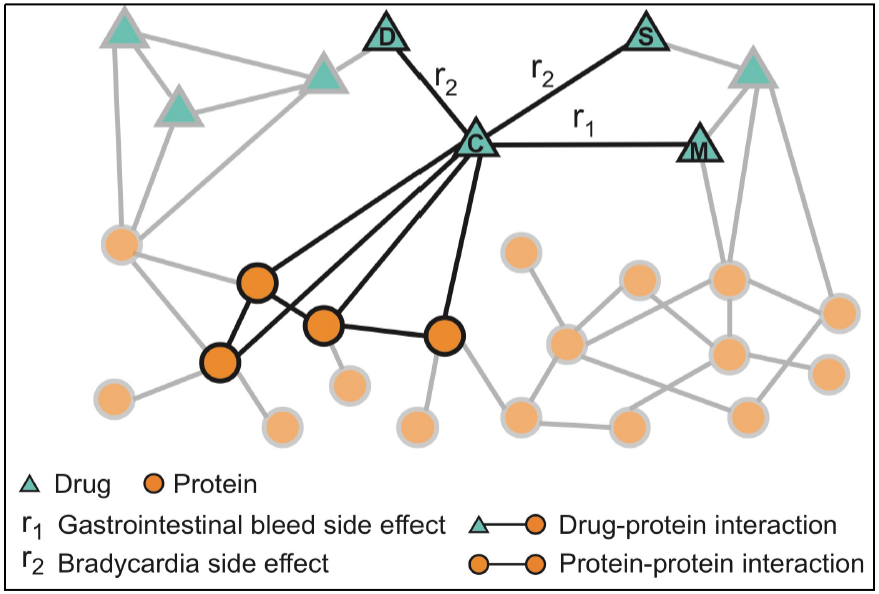
Task: Given a pair of drugs predict adverse side effects



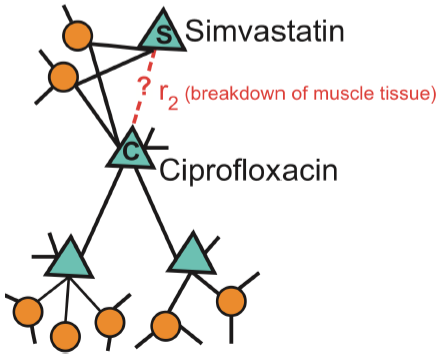
Zitnik et al., Modeling Polypharmacy Side Effects with Graph Convolutional Networks, Bioinformatics 2018

Biomedical Graph Link Prediction

- **Nodes**: Drugs & Proteins
- **Edges**: Interactions



Query: How likely will Simvastatin and Ciprofloxacin, when taken together, break down muscle tissue?



Zitnik et al., Modeling Polypharmacy Side Effects with Graph Convolutional Networks, Bioinformatics 2018

Results: *De novo* Predictions

Rank	Drug c	Drug d	Side effect r	Evidence found
1	Pyrimethamine	Aliskiren	Sarcoma	Stage *et al.* 2015
2	Tigecycline	Bimatoprost	Autonomic neuropathy	
3	Omeprazole	Dacarbazine	Telangiectases	
4	Tolcapone	Pyrimethamine	Breast disorder	Bicker *et al.* 2017
5	Minoxidil	Paricalcitol	Cluster headache	
6	Omeprazole	Amoxicillin	Renal tubular acidosis	Russo *et al.* 2016
7	Anagrelide	Azelaic acid	Cerebral thrombosis	
8	Atorvastatin	Amlodipine	Muscle inflammation	Banakh *et al.* 2017
9	Aliskiren	Tioconazole	Breast inflammation	Parving *et al.* 2012
10	Estradiol	Nadolol	Endometriosis	

Case Report

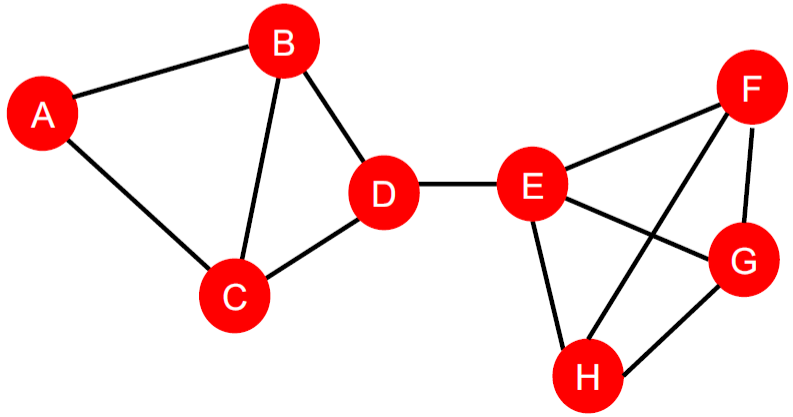
Severe Rhabdomyolysis due to Presumed Drug Interactions between Atorvastatin with Amlodipine and Ticagrelor

Graph Level Tasks

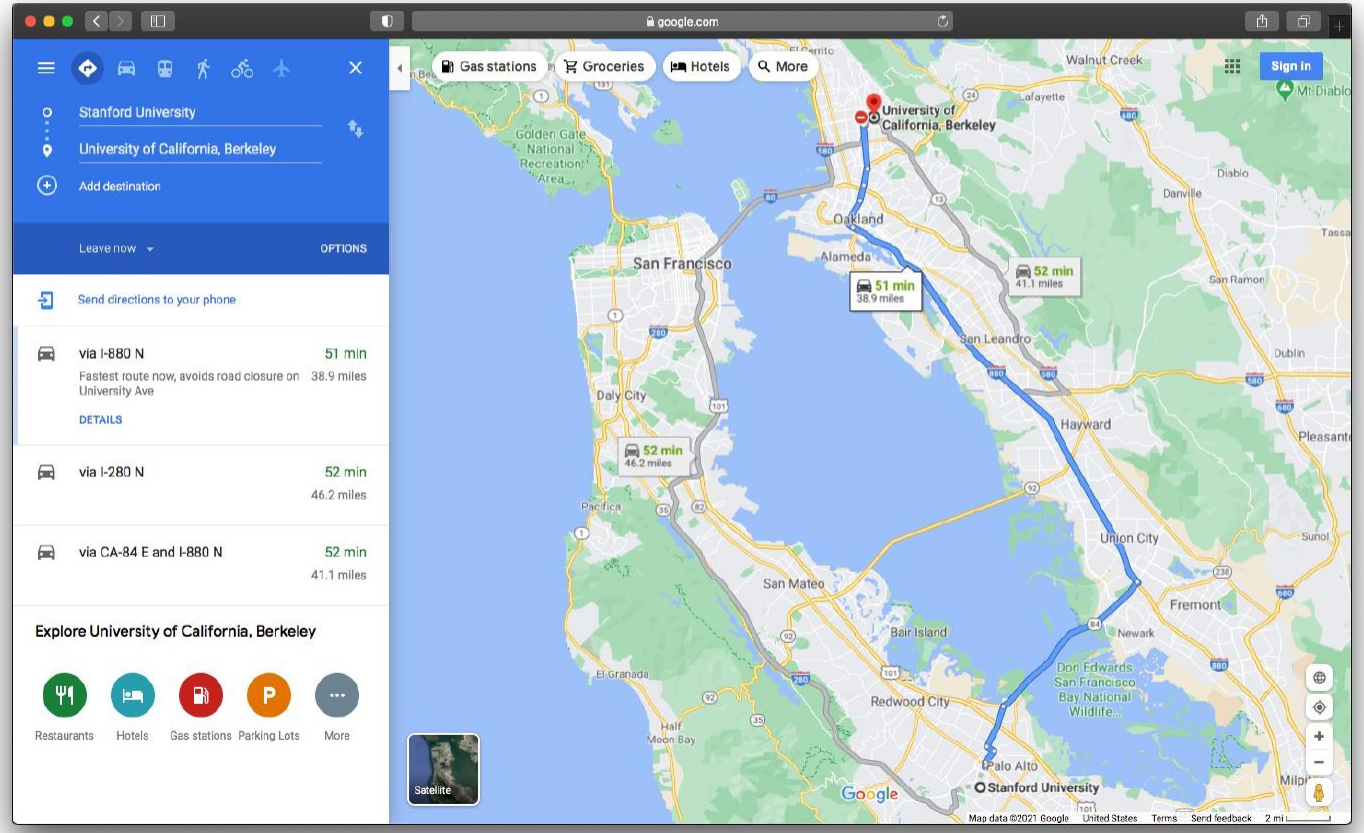
Graph-Level Features

- **Goal:** We want make a prediction for an entire graph or a subgraph of the graph.

- **For example:**



Example (1): Traffic Prediction



Road Network as a Graph

- **Nodes:** Road segments
- **Edges:** Connectivity between road segments
- **Prediction:** Time of Arrival (ETA)

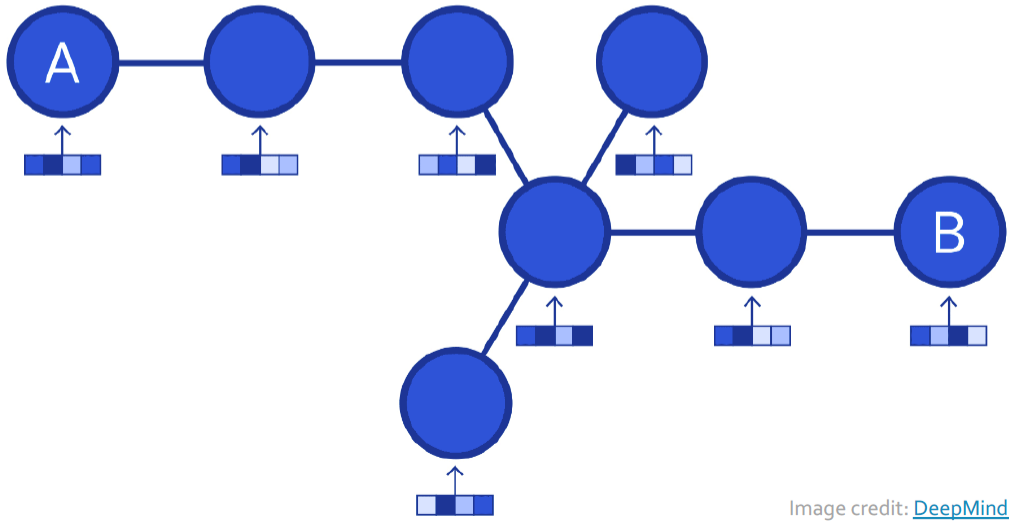
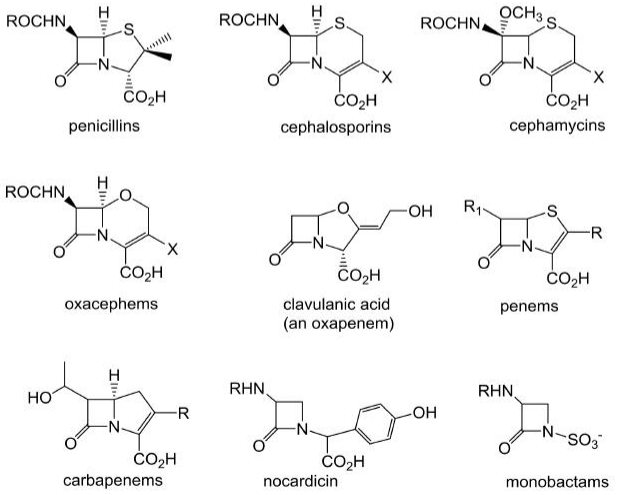


Image credit: DeepMind

Example (2): Drug Discovery

- **Antibiotics are small molecular graphs**
  - **Nodes:** Atoms
  - **Edges:** Chemical bonds



Konaklieva, Monika I. "Molecular targets of β-lactam-based antimicrobials: beyond the usual suspects." Antibiotics 3.2 (2014): 128-142.

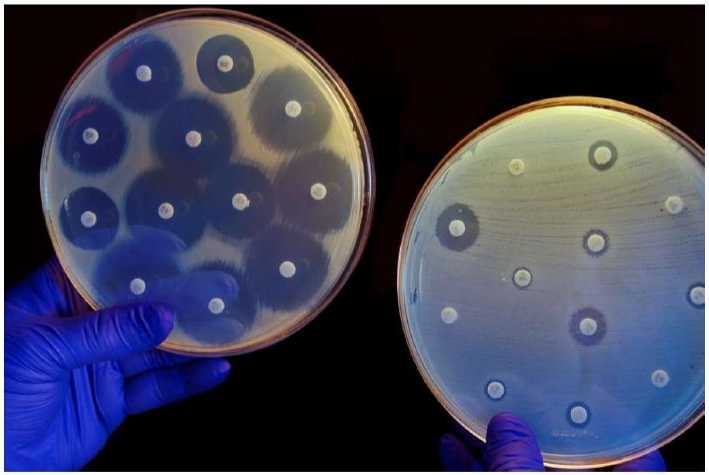


Image credit: CNN

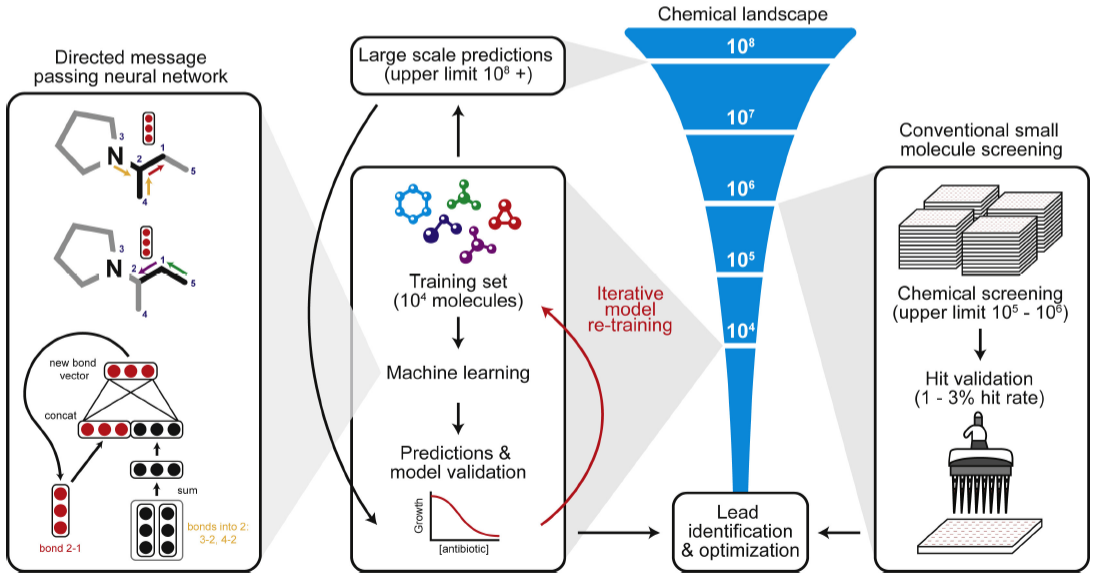
Image credit: DeepMind

THE MODEL ARCHITECTURE FOR DETERMINING OPTIMAL ROUTES AND THEIR TRAVEL TIME.

Stokes et al., A Deep Learning Approach to Antibiotic Discovery, Cell 2020

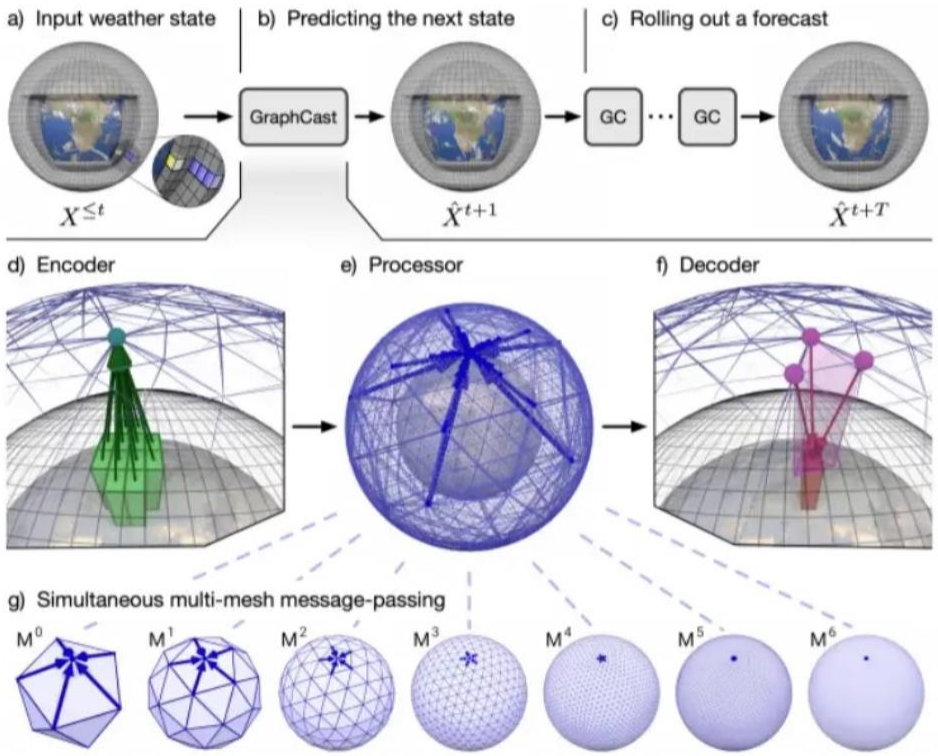
Deep Learning for Antibiotic Discovery

- A Graph Neural Network **graph classification model**
- Predict promising molecules from a pool of candidates



Stokes, Jonathan M., et al. "A deep learning approach to antibiotic discovery." Cell 180.4 (2020): 688-702.

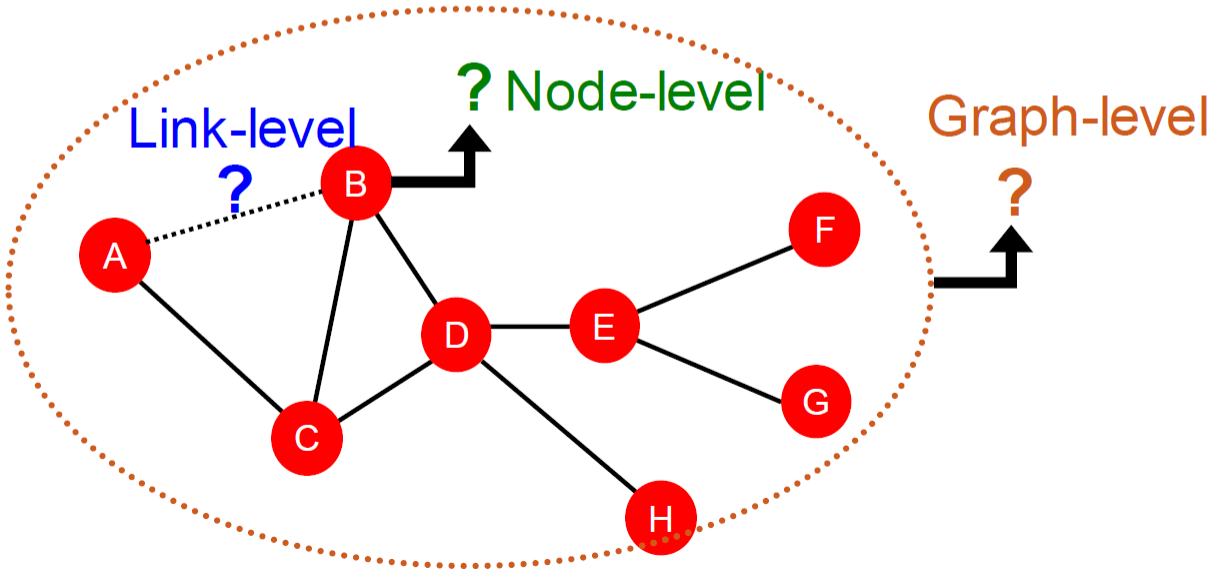
Application: Weather forecasting



https://medium.com/syncedreview/deepmind-googles-ml-based-graphcast-outperforms-the-world-s-best-medium-range-weather-9d114460aa0c

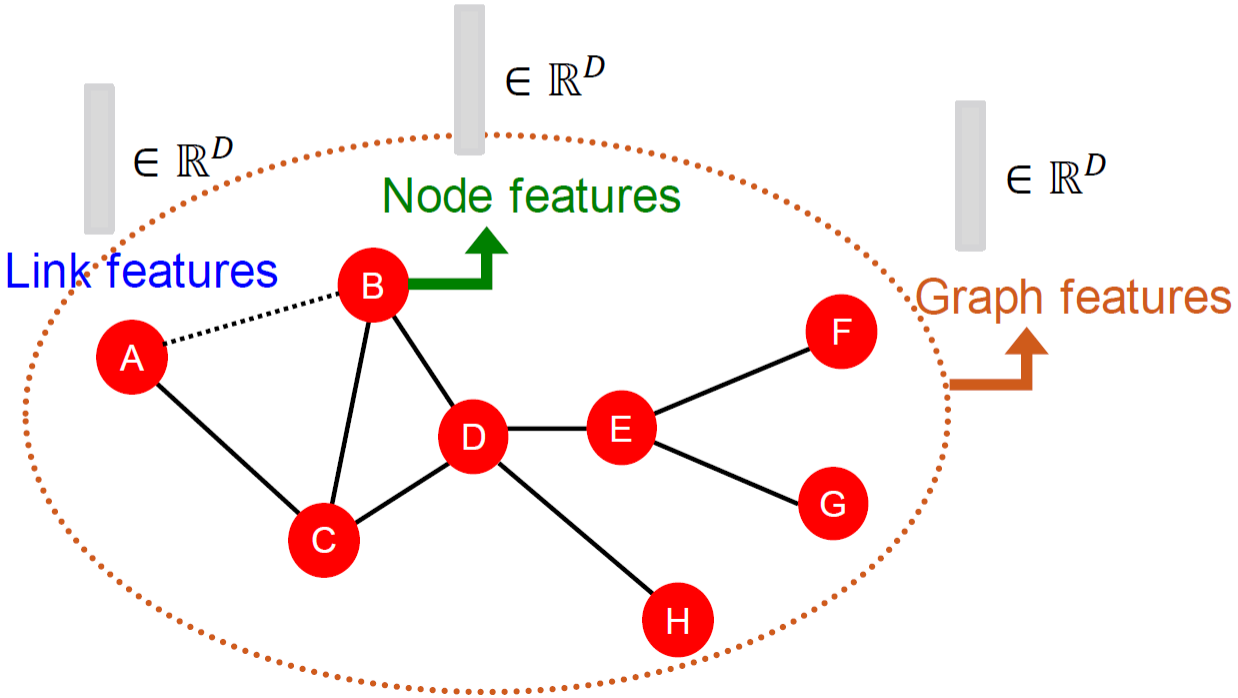
Machine Learning Tasks: Review

- Node-level prediction
- Link-level prediction
- Graph-level prediction



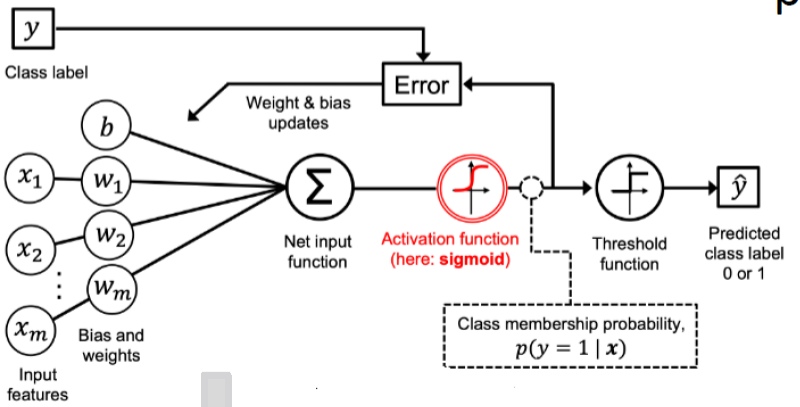
Traditional ML Pipeline

- Design features for nodes/links/graphs
- Obtain features for all training data



Traditional ML Pipeline

- **Train an ML model:**
  - Logistic Regression
  - Random forest
  - Neural network, etc.



$\boldsymbol{x}_N \rightarrow y_N$

- **Apply the model:**
  - Given a new node/link/graph, obtain its features and make a prediction

$\boldsymbol{x} \rightarrow y$

This Lecture: Feature Design

- **Using effective features x over graphs is the key to achieving good model performance.**
- Traditional ML pipeline uses hand-designed features.
- **In this lecture, we overview the traditional features for:**
  - Node-level prediction
  - Link-level prediction
  - Graph-level prediction
- For simplicity, we focus on undirected graphs.

Machine Learning in Graphs

Goal: Make predictions for a set of objects

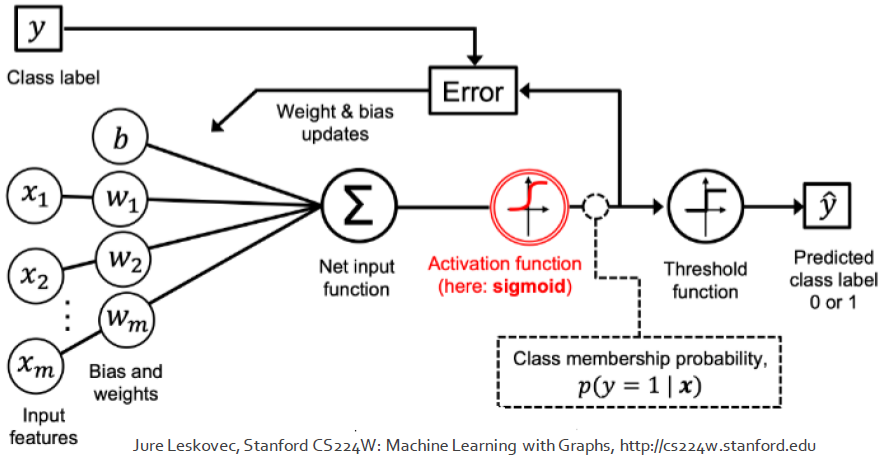
Design choices:

- **Features:** d-dimensional vectors $\boldsymbol{x}$
- **Objects:** Nodes, edges, sets of nodes, entire graphs
- **Objective function:**
  - What task are we aiming to solve?

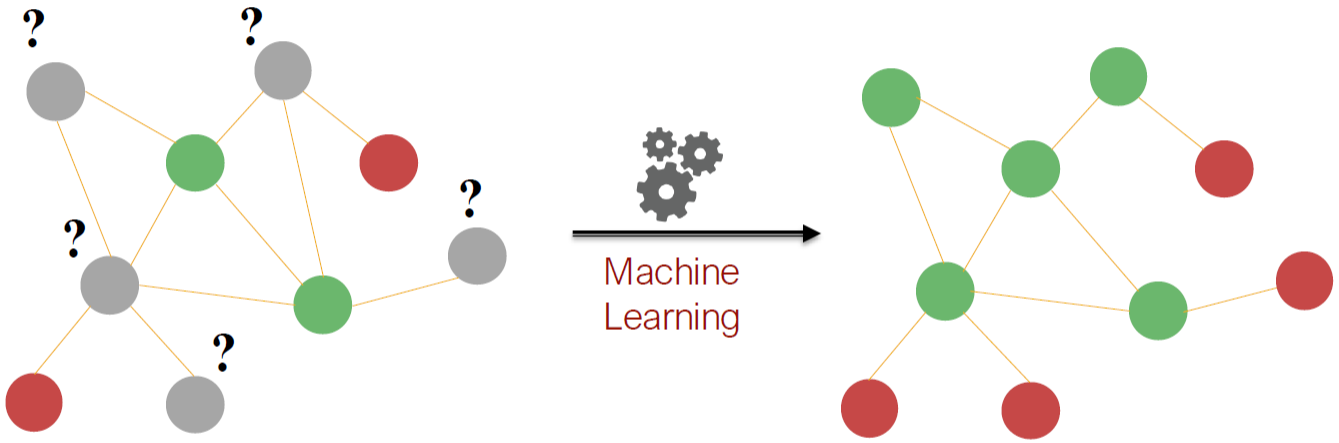
Machine Learning in Graphs

Example: Node-level prediction

- Given: $G = (V, E)$
- Learn a function: $f : V \rightarrow \mathbb{R}$



Node-Level Tasks



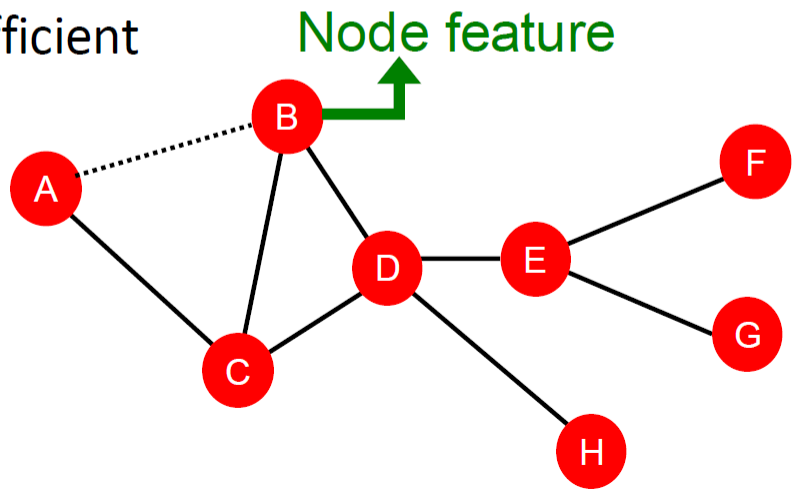
Node classification

ML needs features.

Node-Level Features: Overview

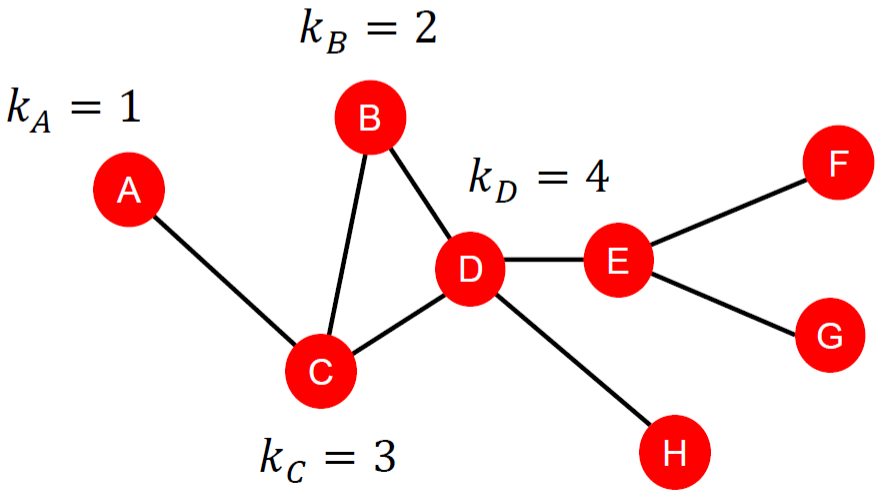
Goal: Characterize the structure and position of a node in the network:

- Node degree
- Node centrality
- Clustering coefficient
- Graphlets



Node Features: Node Degree

- The degree k_v of node v is the number of edges (neighboring nodes) the node has.
- Treats all neighboring nodes equally.



Node Features: Node Centrality

- Node degree counts the neighboring nodes without capturing their importance.
- Node centrality c_v takes the node importance in a graph into account
- **Different ways to model importance:**
  - Eigenvector centrality
  - Betweenness centrality
  - Closeness centrality
  - and many others…

Node Centrality (1)

- **Eigenvector centrality:**
  - A node v is important if surrounded by important neighboring nodes $u \in N(v)$.
  - We model the centrality of node v as the sum of the centrality of neighboring nodes:

$$c_v = \frac{1}{\lambda} \sum_{u \in N(v)} c_u$$

λ is normalization constant (it will turn out to be the largest eigenvalue of A)

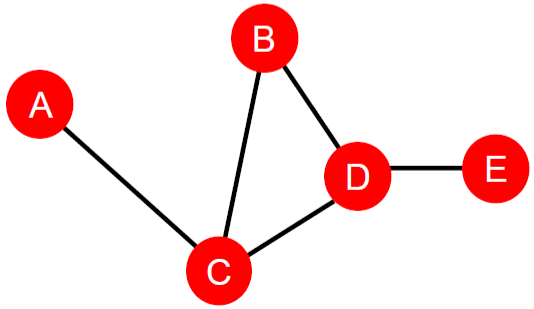
  - Notice that the above equation models centrality in a recursive manner. **How do we solve it?**

Node Centrality (2)

- **Betweenness centrality:**
  - A node is important if it lies on many shortest paths between other nodes.

$$c_v = \sum_{s \neq v \neq t} \frac{\#(\text{shortest paths betwen } s \text{ and } t \text{ that contain } v)}{\#(\text{shortest paths between } s \text{ and } t)}$$

  - **Example:**



$c_A = c_B = c_E = 0$

$c_C = 3$

(A-C-B, A-C-D, A-C-D-E)

$c_D = 3$

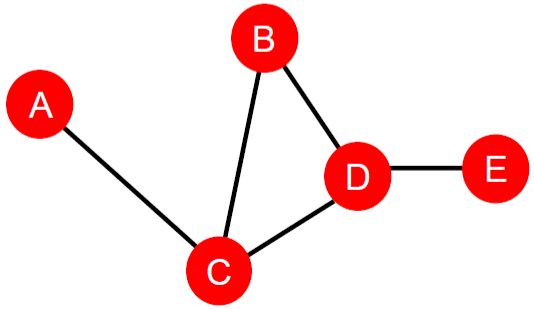
(A-C-D-E, B-D-E, C-D-E)

Node Centrality (3)

- **Closeness centrality:**
  - A node is important if it has small shortest path lengths to all other nodes.

$$c_v = \frac{1}{\sum_{u \neq v} \text{shortest path length between } u \text{ and } v}$$

  - **Example:**



$c_A = 1/(2+1+2+3) = 1/8$
(A-C-B, A-C, A-C-D, A-C-D-E)

$c_D = 1/(2+1+1+1) = 1/5$
(D-C-A, D-B, D-C, D-E)

Node Features: Clustering Coefficient

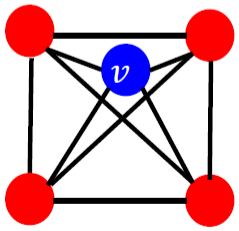
- Measures how connected $v's$ neighboring nodes are:

$$e_v = \frac{\#(\text{edges among neighboring nodes})}{\binom{k_v}{2}} \in [0,1]$$

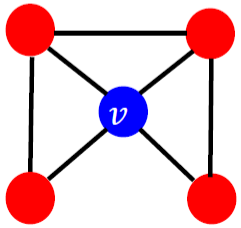
$\#(\text{node pairs among } k_v \text{ neighboring nodes})$

In our examples below the denominator is 6 (4 choose 2).

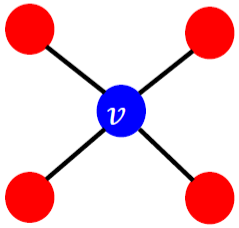
- **Examples:**



$e_v = 1$



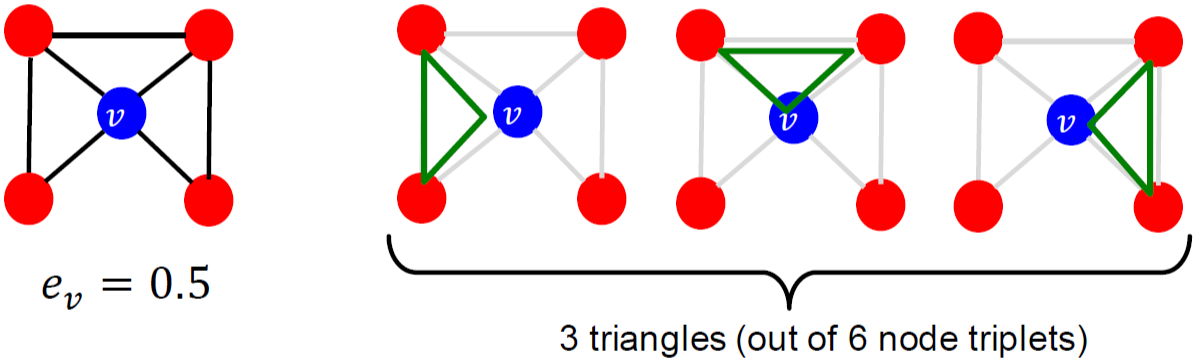
$e_v = 0.5$



$e_v = 0$

Node Features: Graphlets

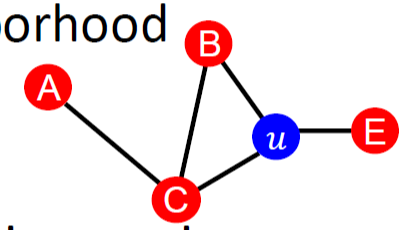
- **Observation:** Clustering coefficient counts the #(triangles) in the ego-network



- We can generalize the above by counting #(pre-specified subgraphs, i.e., **graphlets**).

Node Features: Graphlets

- **Goal:** Describe network structure around node u
  - **Graphlets** are small subgraphs that describe the structure of node u's network neighborhood



Analogy:

- **Degree** counts **#(edges)** that a node touches
- **Clustering coefficient** counts **#(triangles)** that a node touches.
- **Graphlet Degree Vector (GDV):** Graphlet-base features for nodes
  - **GDV** counts **#(graphlets)** that a node touches

Node-Level Feature: Summary

- We have introduced different ways to obtain node features.
- They can be categorized as:
  - Importance-based features:
    - Node degree
    - Different node centrality measures
  - Structure-based features:
    - Node degree
    - Clustering coefficient
    - Graphlet count vector

Node-Level Feature: Summary

- **Importance-based features**: capture the importance of a node in a graph
  - Node degree:
    - Simply counts the number of neighboring nodes
  - Node centrality:
    - Models importance of neighboring nodes in a graph
    - Different modeling choices: eigenvector centrality, betweenness centrality, closeness centrality
- Useful for predicting influential nodes in a graph
  - **Example:** predicting celebrity users in a social network

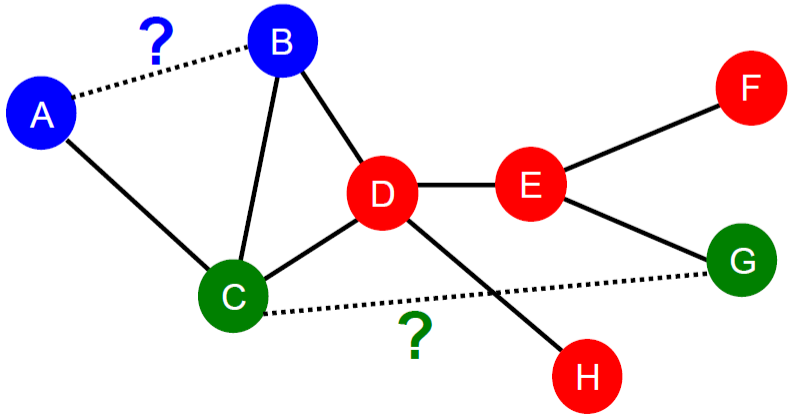
Node-Level Feature: Summary

- **Structure-based features**: Capture topological properties of local neighborhood around a node.
  - **Node degree:**
    - Counts the number of neighboring nodes
  - **Clustering coefficient:**
    - Measures how connected neighboring nodes are
  - **Graphlet degree vector:**
    - Counts the occurrences of different graphlets
- **Useful for predicting a particular role a node plays in a graph:**
  - **Example:** Predicting protein functionality in a protein-protein interaction network.

Link Prediction Task & Features

Link-Level Prediction Task: Recap

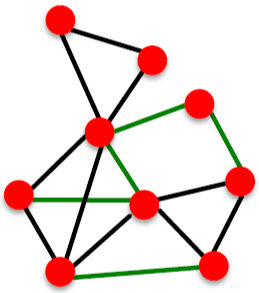
- The task is to predict **new links** based on the existing links.
- At test time, node pairs (with no existing links) are ranked, and top K node pairs are predicted.
- The key is to design features for **a pair of nodes**.



Link Prediction as a Task

Two formulations of the link prediction task:

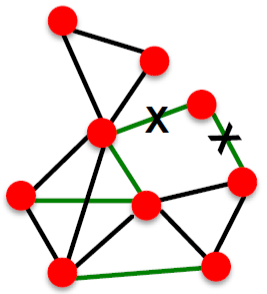
- **1) Links missing at random:**
  - Remove a random set of links and then aim to predict them
- **2) Links over time:**
  - Given $G[t_0, t_0']$ a graph defined by edges up to time t_0', **output a ranked list L** of edges (not in $G[t_0, t_0']$) that are predicted to appear in time $G[t_1, t_1']$
  - **Evaluation:**
    - $n = |E_{new}|$: # new edges that appear during the test period $[t_1, t_1']$
    - Take top n elements of L and count correct edges



$G[t_0, t_0']$
$G[t_1, t_1']$

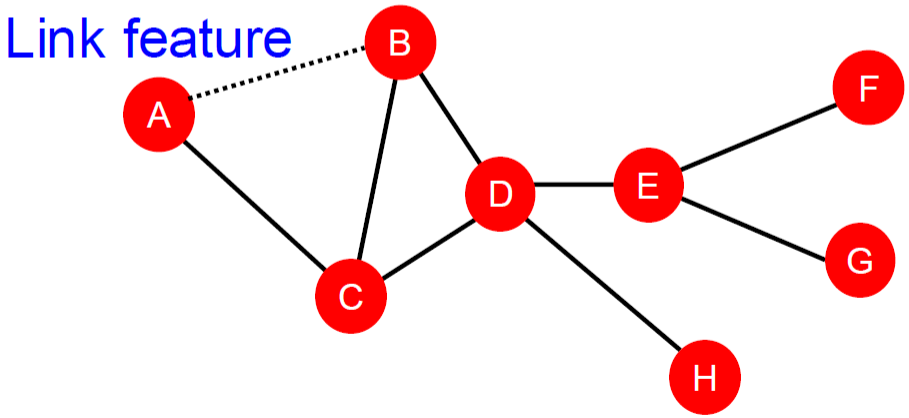
Link Prediction via Proximity

- **Methodology:**
  - For each pair of nodes *(x,y)* compute score *c(x,y)*
    - For example, *c(x,y)* could be the # of common neighbors of *x* and *y*
  - Sort pairs *(x,y)* by the decreasing score *c(x,y)*
  - **Predict top n pairs as new links**
  - **See which of these links actually appear in $G[t_1, t_1']$**



Link-Level Features: Overview

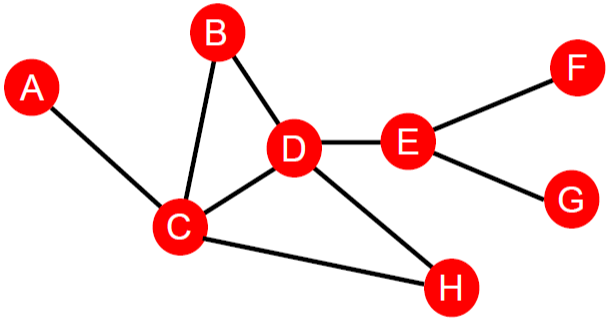
- Distance-based feature
- Local neighborhood overlap
- Global neighborhood overlap



Distance-Based Features

Shortest-path distance between two nodes

- Example:



$$S_{BH} = S_{BE} = S_{AB} = 2$$
$$S_{BG} = S_{BF} = 3$$

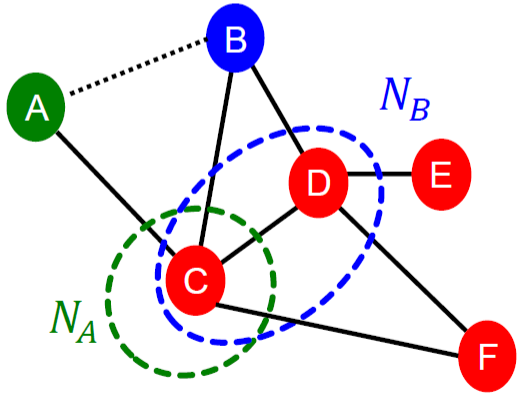
- However, this does not capture the degree of neighborhood overlap:
  - Node pair *(B, H)* has 2 shared neighboring nodes, while pairs *(B, E)* and *(A, B)* only have 1 such node.

Local Neighborhood Overlap

Captures # neighboring nodes shared between two nodes v_1 and v_2:

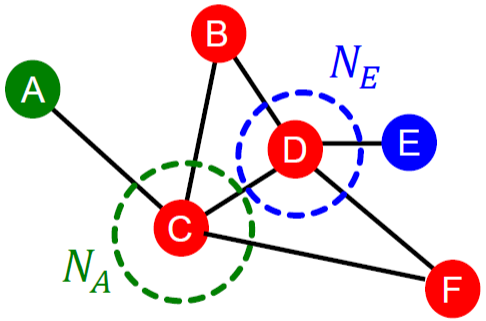
- **Common neighbors:** $|N(v_1) \cap N(v_2)|$
  - Example: $|N(A) \cap N(B)| = |\{C\}| = 1$
- **Jaccard's coefficient:** $\frac{|N(v_1) \cap N(v_2)|}{|N(v_1) \cup N(v_2)|}$
  - Example: $\frac{|N(A) \cap N(B)|}{|N(A) \cup N(B)|} = \frac{|\{C\}|}{|\{C,D\}|} = \frac{1}{2}$
- **Adamic-Adar index:**

  $\sum_{u \in N(v_1) \cap N(v_2)} \frac{1}{\log(k_u)}$
  - Example: $\frac{1}{\log(k_C)} = \frac{1}{\log 4}$



Global Neighborhood Overlap

- **Limitation of local neighborhood features:**
  - Metric is always zero if the two nodes do not have any neighbors in common.



$$N_A \cap N_E = \phi$$
$$|N_A \cap N_E| = 0$$

  - However, the two nodes may still potentially be connected in the future.
- **Global neighborhood overlap** metrics resolve the limitation by considering the entire graph.

Link-Level Features: Summary

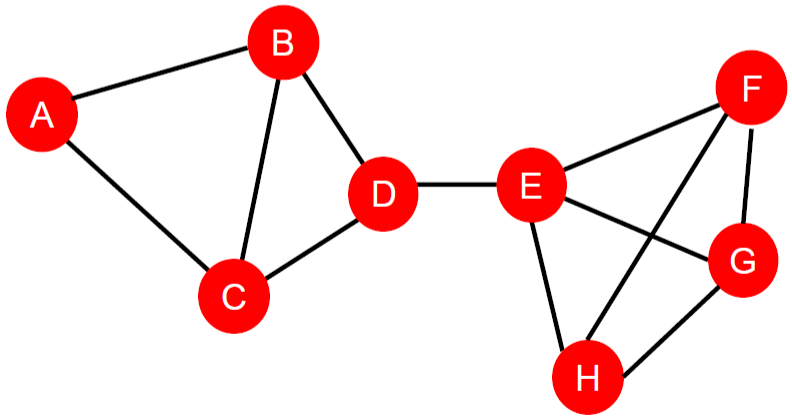
- **Distance-based features:**
  - Uses the shortest path length between two nodes but does not capture how neighborhood overlaps.
- **Local neighborhood overlap:**
  - Captures how many neighboring nodes are shared by two nodes.
  - Becomes zero when no neighbor nodes are shared.
- **Global neighborhood overlap:**
  - Uses global graph structure to score two nodes.
  - Katz index counts #walks of all lengths between two nodes.

Graph Level Features & Graph Kernels

Graph-Level Features

- **Goal:** We want features that characterize the structure of an entire graph.

- **For example:**



Background: Kernel Methods

- **Kernel methods** are widely-used for traditional ML for graph-level prediction.
- **Idea: Design kernels instead of feature vectors.**
- **A quick introduction to Kernels:**
  - Kernel $K(G, G') \in \mathbb{R}$ measures similarity b/w data
  - Kernel matrix $\boldsymbol{K} = \left(K(G, G')\right)_{G,G'}$ must always be positive semidefinite (i.e., has positive eigenvalues)
  - There exists a feature representation $\phi(\cdot)$ such that $K(G, G') = \phi(\mathrm{G})^{\mathrm{T}}\phi(G')$
  - Once the kernel is defined, off-the-shelf ML model, such as kernel SVM, can be used to make predictions.

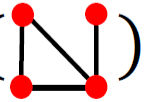
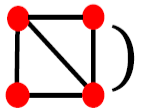
Graph-Level Features: Overview

- **Graph Kernels:** Measure similarity between two graphs:
  - Graphlet Kernel [1]
  - Weisfeiler-Lehman Kernel [2]
  - Other kernels are also proposed in the literature (beyond the scope of this lecture)
    - Random-walk kernel
    - Shortest-path graph kernel
    - And many more…

[1] Shervashidze, Nino, et al. "Efficient graphlet kernels for large graph comparison." Artificial Intelligence and Statistics. 2009.
[2] Shervashidze, Nino, et al. "Weisfeiler-lehman graph kernels." Journal of Machine Learning Research 12.9 (2011).

Graph Kernel: Key Idea

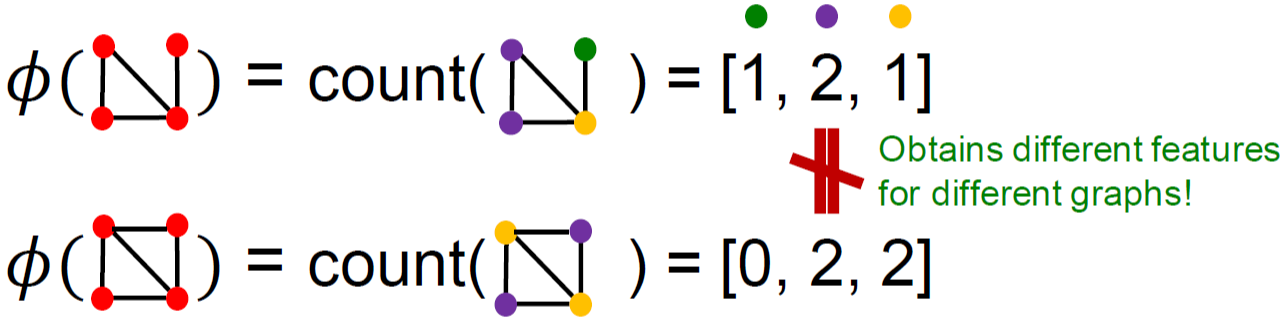
- **Goal**: Design graph feature vector $\phi(G)$
- **Key idea**: Bag-of-Words (BoW) for a graph
  - **Recall:** BoW simply uses the word counts as features for documents (no ordering considered).
  - Naïve extension to a graph: Regard nodes as words.
  - Since both graphs have 4 red nodes, we get the same feature vector for two different graphs…

$\phi($  $) = \phi($  $)$

Graph Kernel: Key Idea

What if we use Bag of **node degrees**?

Deg1: ● Deg2: ● Deg3: ●



- Both Graphlet Kernel and Weisfeiler-Lehman (WL) Kernel use **Bag-of-*** representation of graph, where * is more sophisticated than node degrees!

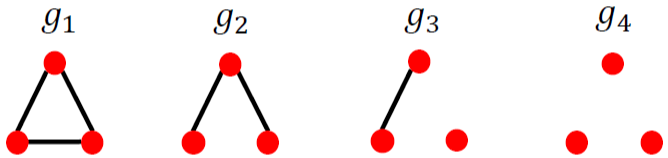
Graph-Level Graphlet Features

- Key idea: Count the number of **different graphlets** in a graph.
  - **Note:** Definition of graphlets here is slightly different from the node-level features.
  - The two differences are:
    - Nodes in graphlets here do **not need to be connected** (allows for isolated nodes)
    - The graphlets here are not rooted.
    - Examples in the next slide illustrate this.

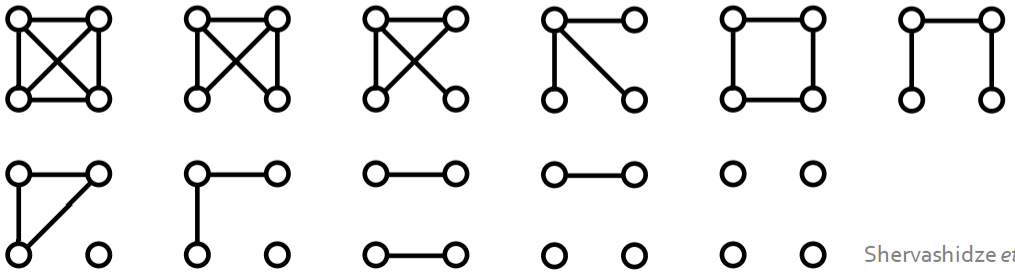
Graph-Level Graphlet Features

Let $G_k = (g_1, g_2, \ldots, g_{n_k})$ be a list of graphlets of size k.

- For $k = 3$, there are 4 graphlets.



- For $k = 4$, there are 11 graphlets.



Shervashidze *et al.*, AISTATS 2011

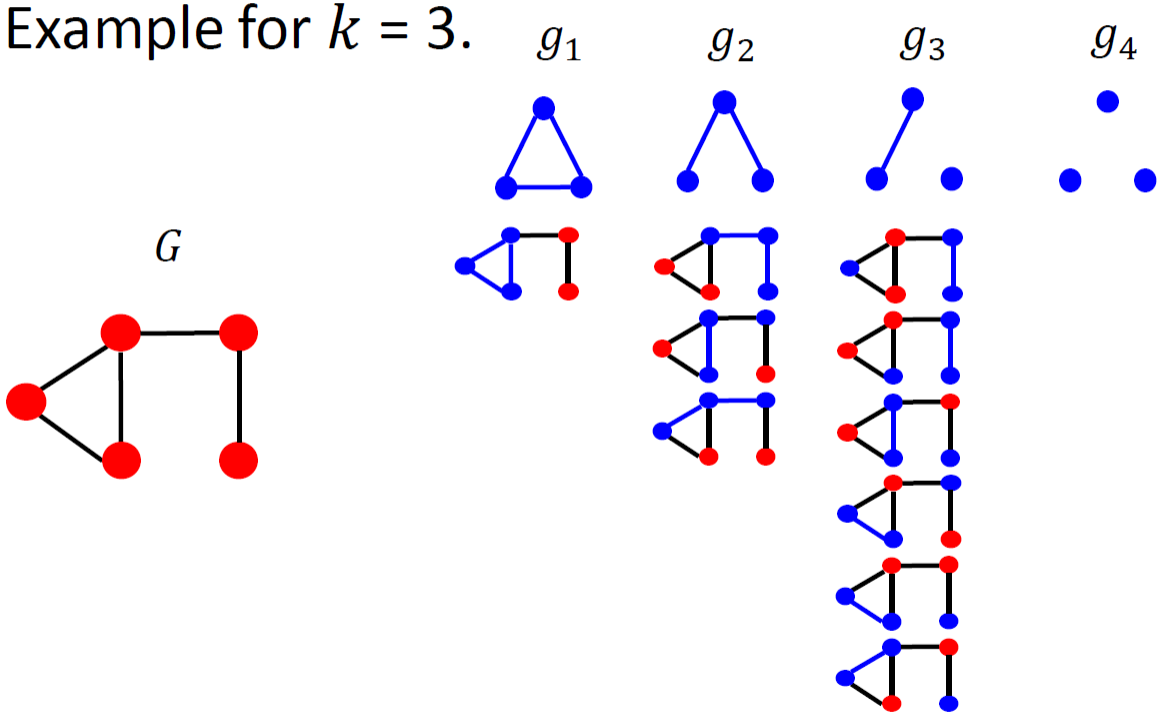
Graph-Level Graphlet Features

- Given graph G, and a graphlet list $\mathcal{G}_k = (g_1, g_2, \ldots, g_{n_k})$, define the graphlet count vector $\boldsymbol{f}_G \in \mathbb{R}^{n_k}$ as

$$(\boldsymbol{f}_G)_i = \#(g_i \subseteq G) \text{ for } i = 1,2,\ldots,n_k.$$

Graph-Level Graphlet Features

- Example for $k = 3$.



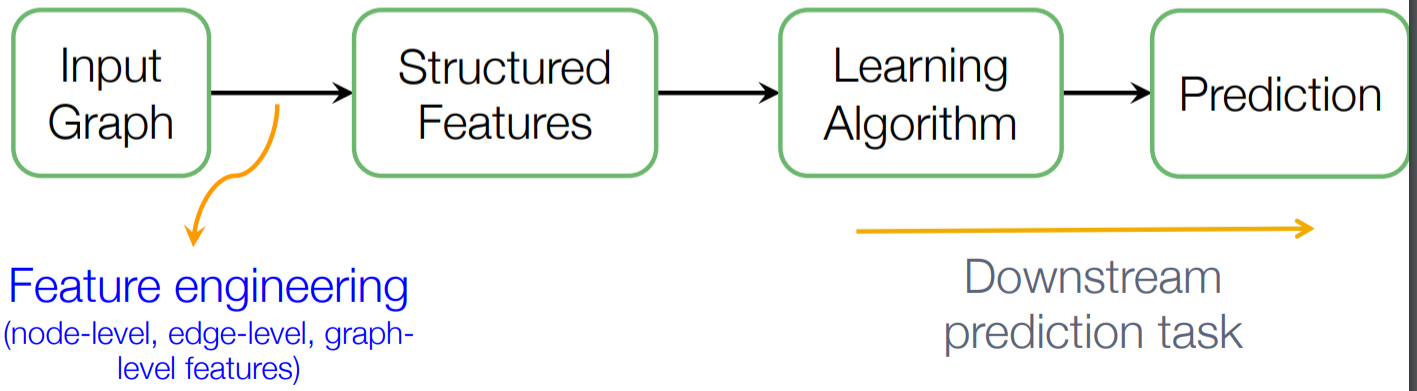
$$f_G = (1, \quad 3, \quad 6, \quad 0)^T$$

Today's Summary

- **Traditional ML Pipeline**
  - Hand-crafted (structural) features + ML model
- **Hand-crafted features for graph data**
  - **Node-level:**
    - Node degree, centrality, clustering coefficient, graphlets
  - **Link-level:**
    - Distance-based feature
    - local/global neighborhood overlap
  - **Graph-level:**
    - Graphlet kernel, WL kernel
- However, we only considered featurizing the graph structure (but not the attribute of nodes and their neighbors)

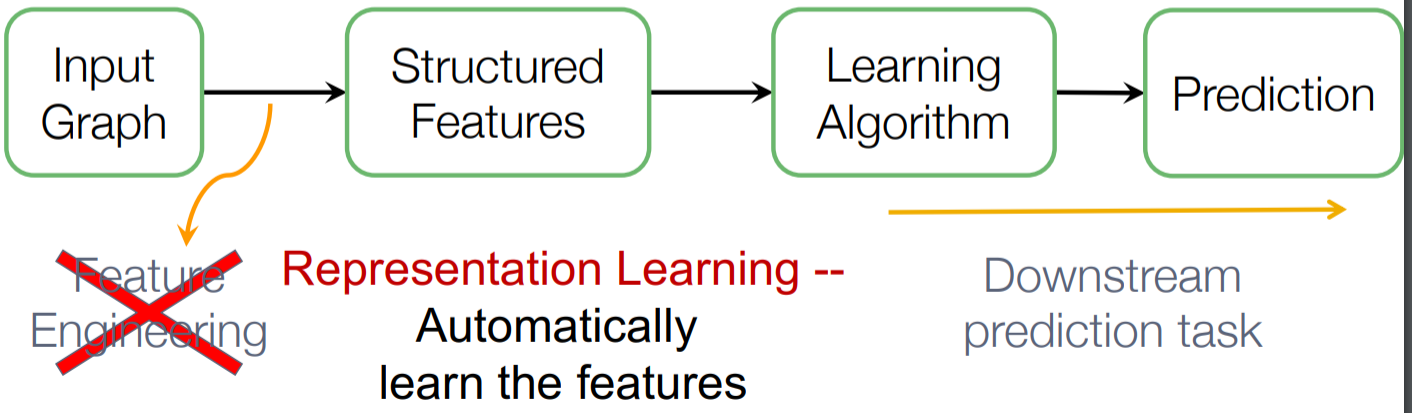
Recap: Traditional ML for Graphs

Given an input graph, extract node, link and graph-level features, then learn a model (SVM, neural network, etc.) that maps features to labels.



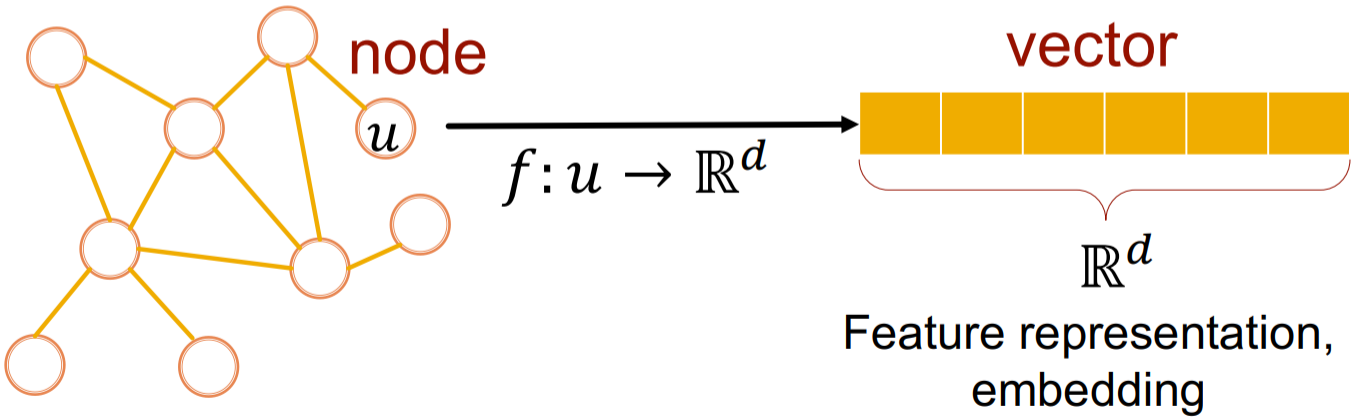
Graph Representation Learning

Graph Representation Learning alleviates the need to do feature engineering every single time.



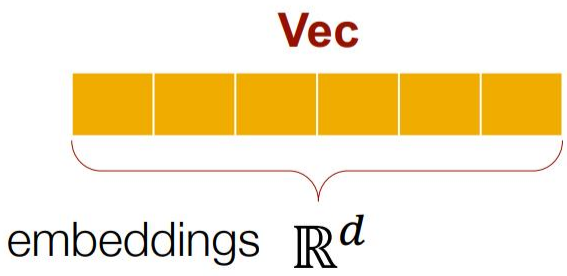
Graph Representation Learning

Goal: Efficient task-independent feature learning for machine learning with graphs!



Why Embedding?

- **Task: Map nodes into an embedding space**
  - Similarity of embeddings between nodes indicates their similarity in the network. For example:
    - Both nodes are close to each other (connected by an edge)
  - Encode network information
  - Potentially used for many downstream predictions



Tasks

- Node classification
- Link prediction
- Graph classification
- Anomalous node detection
- Clustering
- ….

Example Node Embedding

- **2D embedding of nodes of the Zachary's Karate Club network:**

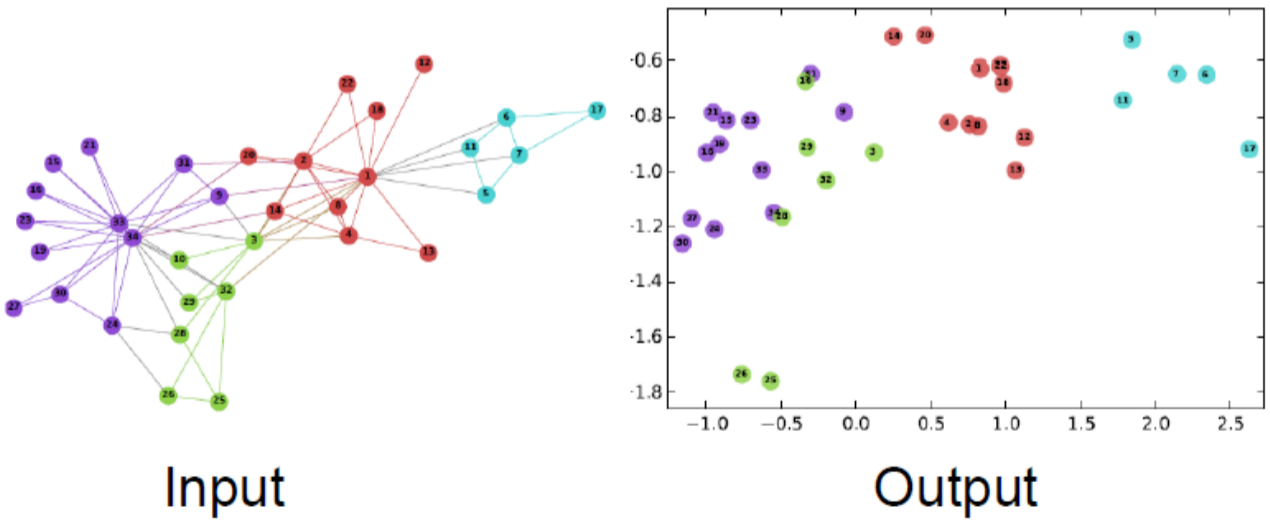
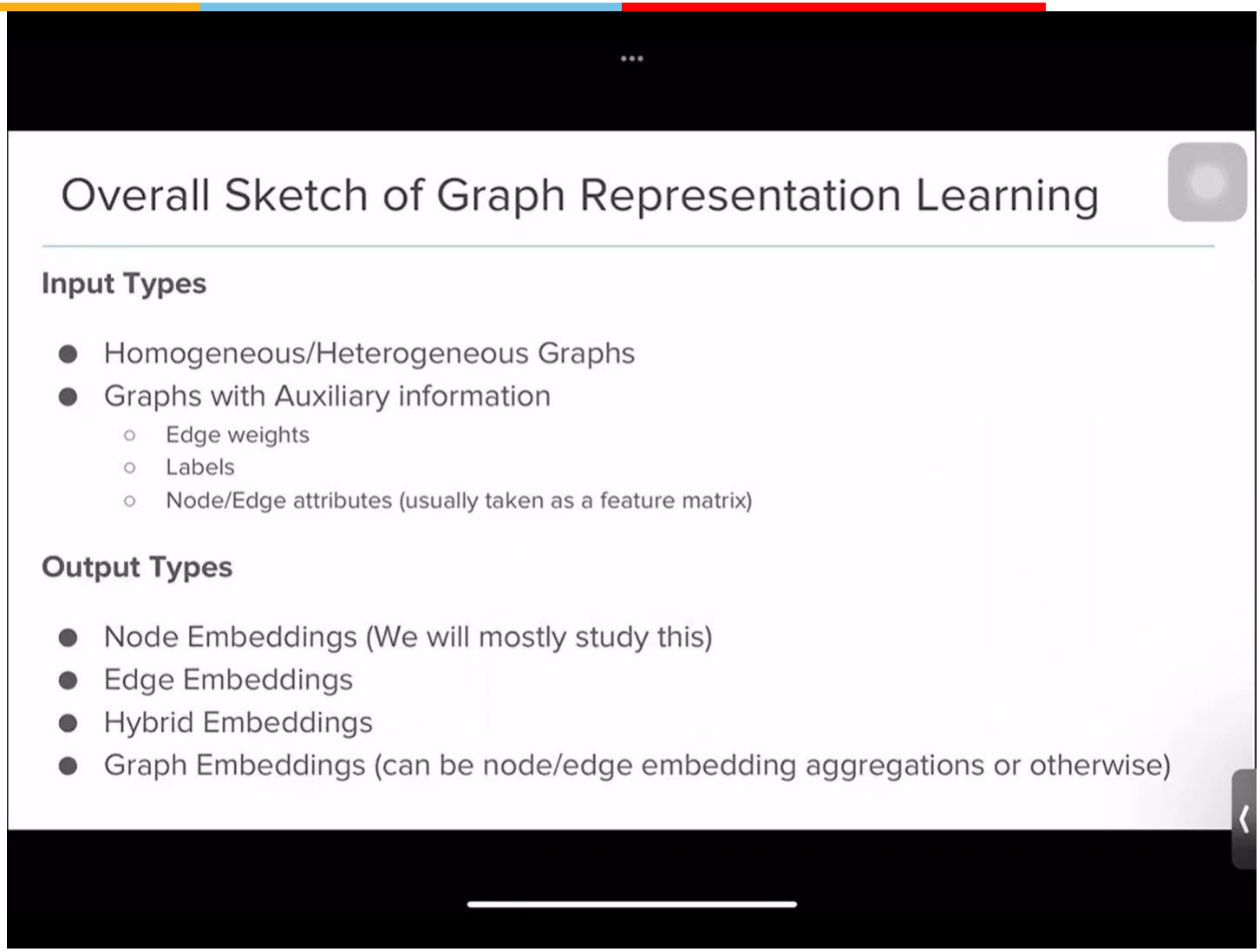


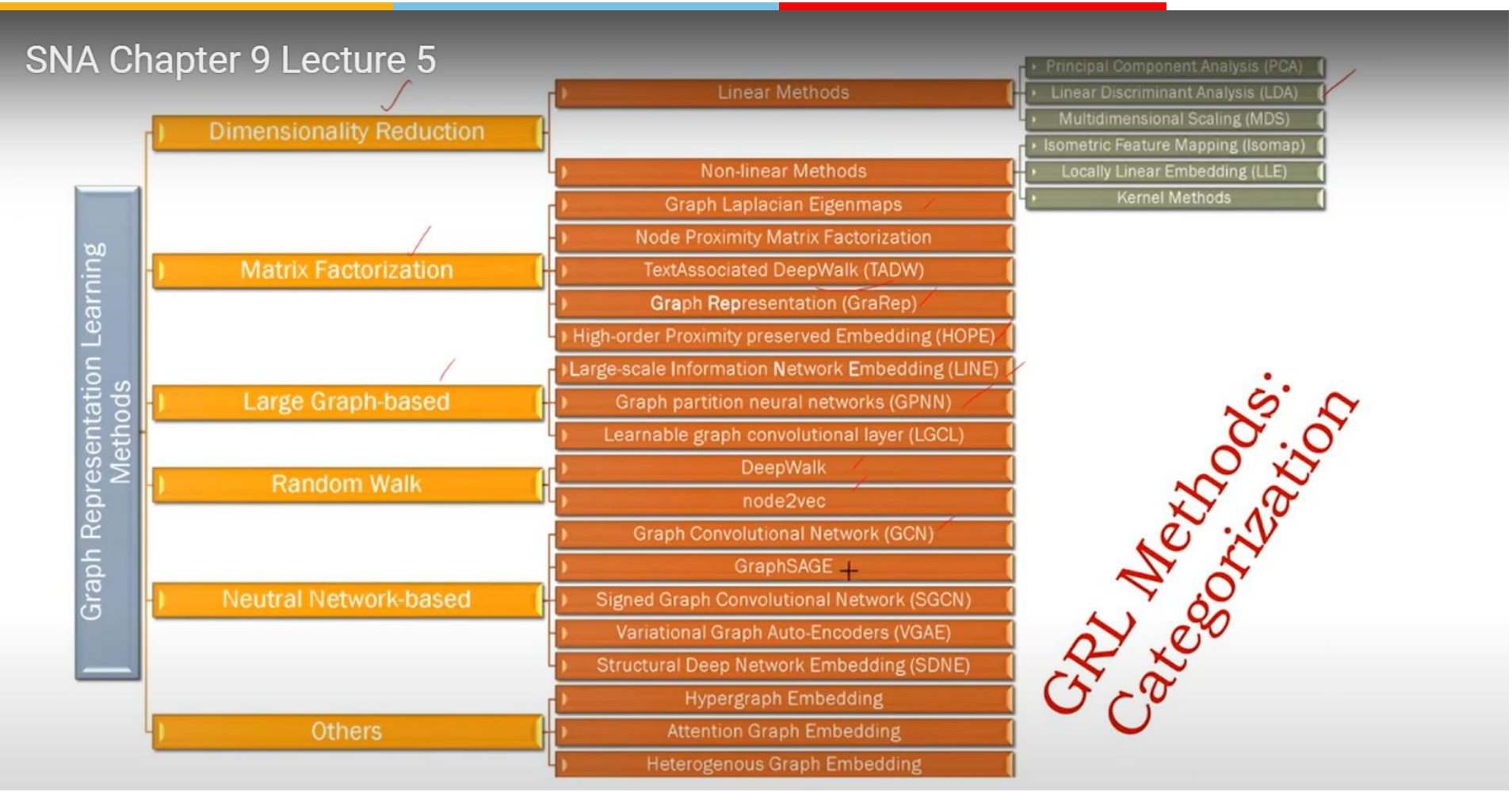
Image from: Perozzi et al. DeepWalk: Online Learning of Social Representations. *KDD 2014.*

Tasks on Networks

Tasks we will be able to solve:

- Node classification
  - Predict the type of a given node
- Link prediction
  - Predict whether two nodes are linked
- Community detection
  - Identify densely linked clusters of nodes
- Network similarity
  - How similar are two (sub)networks


Overall Sketch of Graph Representation Learning
Input Types
Homogeneous/Heterogeneous Graphs
Graphs with Auxiliary information
Edge weights
Labels
Node/Edge attributes (usually taken as a feature matrix)
Output Types
Node Embeddings (We will mostly study this)
Edge Embeddings
Hybrid Embeddings
Graph Embeddings (can be node/edge embedding aggregations or otherwise)



Reference: Tanmoy Chakraborty, https://www.youtube.com/watch?v=NKHgftvRZKM

Demo: Graph Convolution Network

Thank you