COMP4900/5900 Graph Analytics.

Pat Morin 5177 Herzberg Physics Building

Office Hours: see webpage*

Course Webpage(s): Brightspace for submitting assignments, grades, and course forums.

* https://patmorin.me/teaching/graphanal/ for course content.

Grading Scheme: see webpage*

Dates & deadlines: see webpage*

Textbook is freely-available as a PDF on the course webpage (as of January 2026).

About the Course

- Graph mining / analytics
  - subfield of data science studying instances where ~~data items have~~ pairs of data items may or may not have relations between them.
  - Relations are modelled by (directed or undirected) graphs.
  - Examples:
    - social networks with "friend" relation (undirected)
    - microblogging networks with "follows" relation (directed)
    - wireless networks with "within communication range" relation.
    - ...

Notations (as used in class and textbook).

- $\mathbb{N} = \{1, 2, 3, \ldots\}$ is the set of natural numbers $[n] = \{1, 2, \ldots, n\}$
- $\mathbb{Z} = \cancel{\{-2, -1} \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ is the set of integers

$\lfloor x \rfloor$ (floor), $\lceil x \rceil$ (ceiling), $\lfloor x \rceil$ (round) for $z \in \mathbb{Z}$ $\lfloor z + \frac{1}{2} \rceil = z + 1$.

- For a set S
  $|S|$ = the size of S (may be infinite, but is usually countable).

~~For a finit~~

- $P(S)$ = the power set of S (the set of all subsets of S).
- For non-negative integer k, $\binom{S}{k}$ is the set of all k-element subsets of S. $\left|\binom{S}{k}\right| = \binom{|S|}{k}$, where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is a binomial coefficient.
- Notice $2^{|S|} = |P(S)| = \sum_{k=0}^{|S|} \left|\binom{S}{k}\right| = \sum_{k=0}^{|S|} \binom{|S|}{k}$
- Binomial Theorem: $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$

New: For natural numbers $m_1, m_2, \ldots, m_n$ with $d = \sum_{i=1}^{n} m_i$,

$$\binom{d}{m_1, m_2, \ldots m_n} = \frac{d!}{m_1!\, m_2! \cdots m_n!}$$ is called a multinomial coefficient

In COMP2804, these appear when counting rearrangements of strings like SUCCESS $\binom{7}{3, 2, 1, 1} = \binom{7}{3, 2}$ [because SUCCESS has $3 \times S$ and $2 \times C$]

Probability Spaces

$(\Omega, Pr) \leftarrow (\Omega, \mathbb{P})$ in textbook

- Ω is a non-empty countable set.

$Pr : \Omega \to [0,1]$ satisfies $\sum_{\omega \in \Omega} Pr(\omega) = 1$.

Elements of Ω are called outcomes or elementary events.

Subsets of Ω are called events.

For $A \subseteq \Omega$, $Pr(A) \overset{\text{def}}{=} \sum_{\omega \in A} Pr(\omega)$

For $A, B \subseteq \Omega$ with $Pr(B) > 0$,

$Pr(A|B) = \dfrac{Pr(A \cap B)}{Pr(B)}$ is the conditional probability of A given B

Events A and B are independent if $Pr(A \cap B) = Pr(A) \cdot Pr(B)$.
In this case $Pr(A|B) = Pr(A)$.

Events $A_1, \dots, A_n$ are (mutually) independent if for any $I \subseteq [n]$

$$Pr\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} Pr(A_i)$$

Random Variables.

- Any function $X : \Omega = \mathbb{R}$ is called a random variable
- The expected value (or expectation) of X is

$$E(X) \stackrel{\text{def}}{=} \sum_{\omega \in \Omega} X(\omega) \cdot \Pr(\omega)$$

- Random variables X and Y are independent if, for each $x, y \in \mathbb{R}$ the events $\{\omega \in \Omega : X(\omega) = x\}$ and $\{\omega \in \Omega : Y(\omega) = y\}$ are independent. In other words,

$$\Pr(X = x \text{ and } Y = y) = \Pr(X = x) \cdot \Pr(Y = y).$$

- For two random variables X and Y, [Linearity of expectation].

$$E(X+Y) = E(X) + E(Y) \qquad E(a \cdot X + b \cdot Y) = a \cdot E(X) + b \cdot E(Y).$$

New: The variance of X is

$$Var(X) \stackrel{\text{def}}{=} \underbrace{E\left((X - E(X))^2\right) = E(X^2) - (E(X))^2}$$

see next page.

Properties

$$Var(X+b) = Var(X).$$

$$Var(a \cdot X) = a^2 \cdot Var(X)$$

so

$$Var(aX+b) = Var(aX) = a^2 Var(X)$$

For real numbers a, b.

$$\mathrm{Var}(X) = E\left[(X - E(X))^2\right]$$

$$= \sum_{\omega \in \Omega} \Pr(\omega) \cdot \left(X(\omega) - E(X)\right)^2$$

$$= \sum_{\omega \in \Omega} \Pr(\omega) \cdot \left((X(\omega))^2 - 2 \cdot X(\omega) \cdot E(X) + (E(X))^2 \right)$$

$$= \sum_{\omega \in \Omega} \Pr(\omega) \cdot (X(\omega))^2 - \sum_{\omega} 2 \cdot X(\omega) \cdot E(X) + \sum_{\omega} (E(X))^2$$

$$= \quad E(X^2) \quad - 2 \cdot E(X) \cdot \sum_{\omega} X(\omega) \quad + (E(X))^2$$

$$= \quad E(X^2) \quad - 2(E(X))^2 \qquad + (E(X))^2$$

$$= E(X^2) - (E(X))^2 .$$

$$\mathrm{Var}(X+b) = E\left((X+b)^2\right) - (E(X+b))^2 \quad = E(X^2) + 2 \cdot E(bX) + E(b^2) - (E(X)+b)^2$$

$$= E(X^2) + 2 \cdot b \cdot E(X) + b^2 \quad - (E(X))^2 - 2 \cdot b \cdot E(X) - b^2$$

$$= E(X^2) - (E(X))^2$$

$$= \mathrm{Var}(X).$$

$$\mathrm{Var}(aX) = E\left((aX)^2\right) - (E(aX))^2$$

$$= a^2 \cdot E(X^2) \quad - a^2 (E(X))^2$$

$$= a^2 \cdot \left(E(X^2) - (E(X))^2\right) = a^2 \, \mathrm{Var}(X).$$

For two random variables X and Y, the covariance is

$$\underline{Cov(X,Y)} \stackrel{def}{=} E\big((X-E(X))\cdot(Y-E(Y))\big) = E(X\cdot Y) - E(X)\cdot E(Y).$$

Note that $Cov(X,X) = E\big((X-E(X))^2\big) = Var(X)$.

$$\begin{aligned} Var(X+Y) &= \cancel{E\big((X+Y)^2 - E(X+Y)\big)} \\ & E\big((X+Y)^2\big) - \big(E(X+Y)\big)^2 \\ &= E(X^2) + 2\cdot E(XY) + E(Y^2) - (E(X))^2 - E(X)\cdot E(Y) - (E(Y))^2 \\ &= Var(X) + 2Cov(X,Y) + Var(Y). \end{aligned}$$

Note: If X and Y are independent, then

$$E(X\cdot Y) = E(X)\cdot E(Y)$$

$$E(X-E(X)) = E(X) - E(X) = 0.$$

So, if X and Y are independent then

$$Cov(X,Y) = E\big((X-E(X))\cdot(Y-E(Y))\big) = E(X-E(X))\cdot E(Y-E(Y)) = 0\cdot 0 = 0.$$

Roughly: $Cov(X,Y)$ measures how dependent X and Y are.

Pearson's Correlation Coefficient normalizes this into $[-1,1]$:

$$\rho_{X,Y} \stackrel{def}{=} \frac{Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}$$

An <u>undirected graph</u> $G=(V,E)$ has a set V of <u>vertices</u> (or <u>nodes</u>) and a set $E \subseteq \binom{V}{2}$ of <u>edges</u>

We write $\{v,w\} \in E$ as vw, so there is no difference between vw and wv (both refer to the 2-element set $\{v,w\}$).

A <u>directed graph</u> $D=(V,E)$ has a set V of <u>vertices</u>/<u>nodes</u>, but E is a set of ordered pairs ~~all elements~~ of vertices.

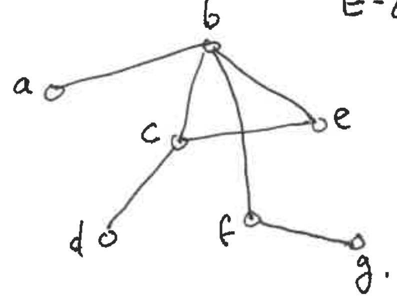
We write $(v,w) \in E$ as vw (or sometimes $\vec{vw}$). Now there is a difference between $vw=(v,w)$ and $wv=(w,v)$ (the same pair but ordered differently).

The vertices v and w of an edge vw are the <u>endpoints</u> of the edge vw. If vw is an edge of a directed graph then v is the <u>source</u> of the edge vw and w is the <u>target</u> of the edge vw.

Drawing Graphs.

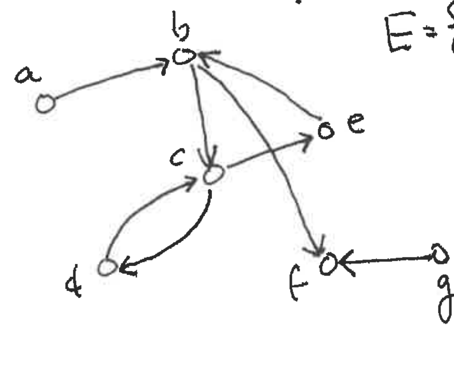
Undirected Graph.

$E=\{ab, bc, ce, eb, cd, bf, fg\}$.



Directed Graph

$E=\{ab, bc, ce, eb, cd, dc, bf, gf\}$.



Adjacency Matrix.

- Represents a graph by a $|V| \times |V|$ binary matrix M

$$M_{vw} = \begin{cases} 1 & \text{if } vw \in E \\ 0 & \text{otherwise.} \end{cases}$$

- If G is undirected, then M is symmetric $M_{vw} = M_{wv}$

Other representations are possible (and usually preferable) including edge lists, adjacency lists, or incidence matrices.

Weighted Graphs.

- In some applications, the edges of a graph have (real-valued) weights.

Paths & Connected Components.

- A path is a sequence of vertices $v_0, \ldots, v_\ell$ such that $v_{i-1} v_i \in E$ for each $i \in \{1, \ldots, \ell\}$.

- Two vertices v and w in an undirected graph $G = (V, E)$ are in the same connected component of G if and only if G contains a path $v_0, \ldots, v_\ell$ with $v = v_0$ and $w = v_\ell$.

~~Two vertices~~

A graph with only one connected component that contains all vertices is a connected graph.
undirected

The distance between two vertices v and w in the same connected component is the minimum number of edges in a path with endpoints v and w. [So if $v_0, \ldots, v_\ell$ is a path from v to w then $\mathrm{dist}(v,w) \leq \ell$.].

Degree Distribution

$$n_\ell = |\{v \in V : \deg(v) = \ell\}|$$

$$d_\ell = \frac{n_\ell}{|V|}$$

Define probability space (Ω, Pr) with $\Omega = V$ and $Pr(\omega) = \frac{1}{|V|}$ for each $\omega \in \Omega$. Then define random variable X by

$$X(\omega) = \deg(\omega) \text{ for each } \omega \in \Omega.$$

Then $d_\ell = Pr(X = \ell)$

$$E(X) = \frac{1}{|V|} \cdot \sum_{\omega \in V} \deg(\omega) = \frac{2 \cdot |E|}{|V|} = \text{average degree of } G.$$

Any statistic on the degree-distribution $d_0, d_1, d_2, \ldots$ can be treated as a statistic about the graph $G = (V, E)$.

For example if $Var(X) = 0$ then every vertex of G has degree $E(X)$. [This is called a regular graph]

There are many measures of distance between two distributions These can be used to compare two graphs G_1 and G_2 even if they have different numbers of vertices.

Digression: Induced Subgraphs and Notation.

Sometimes we are studying more than one graph at the same time, and sometimes two graphs we are studying even share vertices, so notations like $N(v)$ and $\deg(v)$ become ambiguous.

More: If we're studying two graphs G and H, then what do V and E refer to?

Don't confuse with expected value. We'll (try to) use $\mathbb{E}(X)$ for that

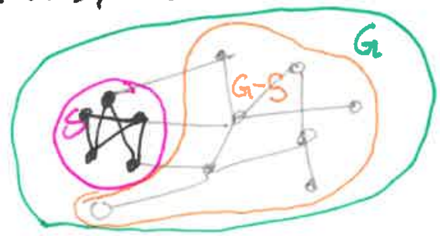
Fix: For any graph G we will use $V(G)$ to refer to the vertex ~~edge~~ set of G and $E(G)$ to refer ~~to~~ the edge set of G.

When there is any danger of confusion, we will use the notations

- $n_G = |V(G)|$ $m_G = |E(G)|$
- $\deg_G(v)$ = the degree of v in the graph G
- $N_G(v)$ = the neighbours of v in G (Notice $\deg_G(v) = |N_G(v)|$)
- $\mathrm{dist}_G(v,w)$ = the distance between v and w in G.
- and so on....

Induced Subgraphs

For a graph G and $S \subseteq V(G)$, the subgraph of G induced by S is the graph $G[S]$ with vertex set $V(G[S]) = S$ and edge set $E(G[S]) = \{vw \in E(G) : v, w \in S\}$.



Also $G - S = G[V(G) \setminus S]$

Average Degree versus Edge Density.

If G has n vertices and m edges, then the average degree of G is

$$d(G) = \frac{1}{n}\sum_{v \in V(G)} \deg_G(v) = \frac{2m}{n}.$$

<u>Notice:</u> $0 \leq d(G) \leq n-1$

The edge density of G is

$$\text{density}(G) = \frac{m}{\binom{n}{2}}$$

measures the fraction of $\binom{n}{2}$ possible edges that are actually in G.

density(G) is also a probability:

- choose 2 random vertices ~~of~~ $\{v,w\} \in \binom{V(G)}{2}$.
- $\text{density}(G) = \Pr(vw \in E(G))$.

Clustering Coefficient.

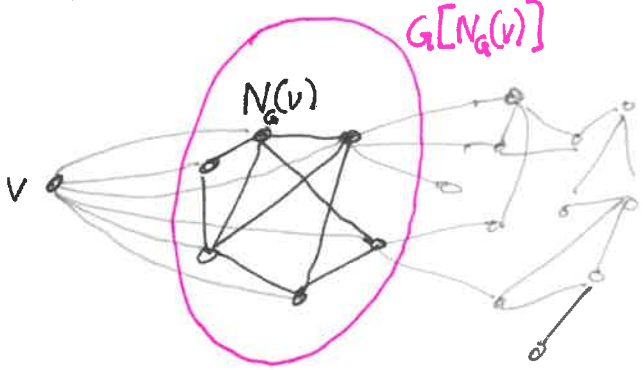
For a vertex v in a graph G, the local clustering coefficient (at v) is

$$c_G(v) \overset{\text{def}}{=} \frac{|E(G[N_G(v)])|}{\binom{\deg_G(v)}{2}} = \frac{|\{uw \in E(G) : u,w \in N_G(v)\}|}{\binom{\deg_G(v)}{2}}.$$

This measures the edge density of $G[N_G(v)]$. Since $G[N_G(v)]$ has ~~at most~~ $\deg_G(v)$ vertices, it has at most $\binom{\deg_G(v)}{2}$ edges, so $0 \le c(v) \le 1$ for each vertex v of G.

<u>Note:</u> $c_G(v)$ is not defined if $\deg_G(v) < 2$, since then $\binom{\deg_G(v)}{2} = 0$, so the definition of $c_G(v)$ divides by zero in this case.

For any $d \geq 2$, $C_G(d) = \dfrac{\sum_{\{v \in V(G) : \deg_G(v) = d\}} c_G(v)}{|\{v \in V(G) : \deg_G(v) = d\}|}$ is the average clustering coefficient over all vertices of degree d. The $C_G(2), C_G(3), C_G(4), \ldots\ldots$ is a sequence of real numbers that characterizes G.



When is $C(d)$ considered large?

If G has n vertices and m edges and we choose $S \in \binom{V(G)}{d}$ uniformly at random, then for any fixed edge $e \in E(G)$,

$$\Pr(e \in E(G[S])) = \frac{\binom{n-2}{d-2}}{\binom{n}{d}} = \frac{\left(\frac{(n-2)!}{(d-2)!(n-d)!}\right)}{\left(\frac{n!}{d!(n-d)!}\right)} = \frac{d(d-1)}{n(n-1)}$$

~~Define~~ For each $e \in E(G)$, define

$$X_e = \begin{cases} 1 & \text{if } e \in E(G[S]) \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathbb{E}(X_e) = \Pr(X_e = 1) = \dfrac{d(d-1)}{n(n-1)}$.

Let $X = \sum_{e \in E(G)} X_e$ count the number of edges in $G[S]$.

Then $\mathbb{E}(|E(G[S])|) = \mathbb{E}(X) = \mathbb{E}\left(\sum_{e \in E(G)} X_e\right) = \sum_{e \in E(G)} \mathbb{E}(X_e) = \dfrac{m \cdot d \cdot (d-1)}{n \cdot (n-1)}$.

And $\mathbb{E}\left(\dfrac{|E(G[S])|}{\binom{d}{2}}\right) = \dfrac{\mathbb{E}(X)}{\binom{d}{2}} = \dfrac{2m}{n(n-1)} = \dfrac{m}{\binom{n}{2}} =$ the edge density of G.

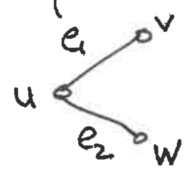
So, if $C(d)$ is much bigger than $\frac{m}{\binom{n}{2}}$ then this indicates that the neighbourhoods of degree-d vertices have more edges than random subsets of d vertices.

The friends of v are more likely to be friendly with each other than a random group of people.

Global Clustering Coefficient

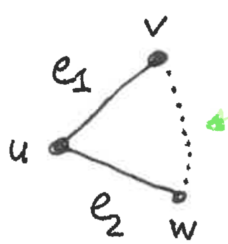
$$C_{glob}(G) = \frac{(\#\text{ 3-cycles in } G)\cdot 3}{\sum_{v\in V(G)} \binom{\deg_G(v)}{2}}$$

The denominator in $C_{glob}(G)$ counts the number of pairs of edges $e_1, e_2 \in E(G)$ that have a common vertex:



~~divided by 3~~.

$C_{glob}(G)$ is the probability that, if we pick a random pair of adjacent edges e_1 and e_2, that these edges are part of a 3-cycle.



$\Pr(vw \in E(G)) = C_{glob}(G)$

On the other hand, if we pick two random vertices $\{v,w\} \in \binom{V(G)}{2}$,

$$\Pr(vw \in E(G)) = \frac{m}{\binom{n}{2}}.$$

So, if $C_{glob}(G) \gg \frac{m}{\binom{n}{2}}$, then this ~~indicates~~ is an indication that edges cluster together.

Local Clustering Coefficient

$$C_{loc}(G) = \frac{\sum_{v \in V(G) : \deg_G(v) \geq 2} c(v)}{|\{v \in V(G) : \deg_G(v) \geq 2\}|}$$

is the averages of the clustering coefficients of vertices of G (at least those vertices where $c(v)$ is defined).

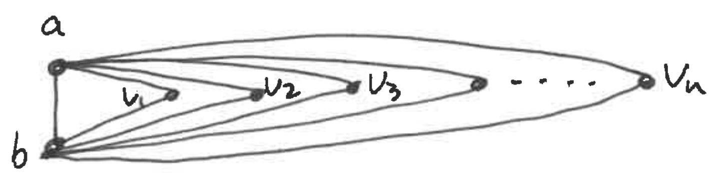
$C_{loc}(G)$ is also the probability that some edge exists:

- Pick a random degree ≥ 2 vertex $u \in V(G)$.
- Pick two random vertices $\{v, w\} \in \binom{N_G(u)}{2}$.

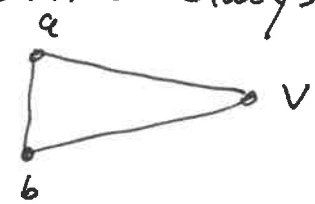
Then $C_{loc}(G)$ is the probability that $vw \in E(G)$.

$C_{loc}(G)$ and $C_{glob}(G)$ can be very different.

Bike wheel graph B_n



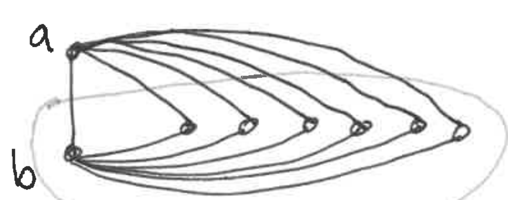
- Pick a random vertex. Almost always a vertex of degree 2 with $c(v) = \frac{1}{\binom{2}{2}} = 1$



so $C_{loc}(B_n) \approx 1$.

- There are $\binom{n}{2}$ pairs of ~~incident~~ edges incident to a and $\binom{n}{2}$ pairs of edges incident to b, but only n pairs of edges incident to $v_1, \ldots, v_n$.

- Pick a random pair of incident edges. Almost always a pair incident to a or a pair incident to b.

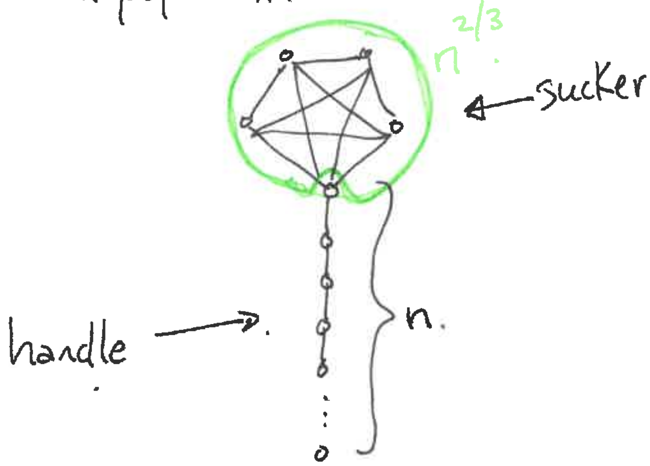


$N_G(a)$ has $n+1$ vertices and only n edges, so

$$c_{B_n}(a) = \frac{n}{\binom{n+1}{2}} = \frac{2}{n+1} \approx 0.$$

So $C_{glob}(G) \approx 0$.

Lollipop L_n.



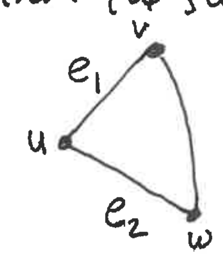
Random vertex is almost always part of the handle and has $\deg_{L_n}(v) = 2$

$$c_{L_n}(v) = \frac{0}{\binom{2}{2}} = 0$$

so $C_{loc}(L_n) \approx 0$.

Handle only has $O(n)$ pairs of incident edges, but the sucker has $\Theta(n^{4/3})$ pairs of incident edges.

Random pair of incident edges is almost always from the sucker.
and ~~its neighbourhood is a clique, s~~ the three vertices form a triangle.



So $C_{glob}(L_n) \gtrsim 1$