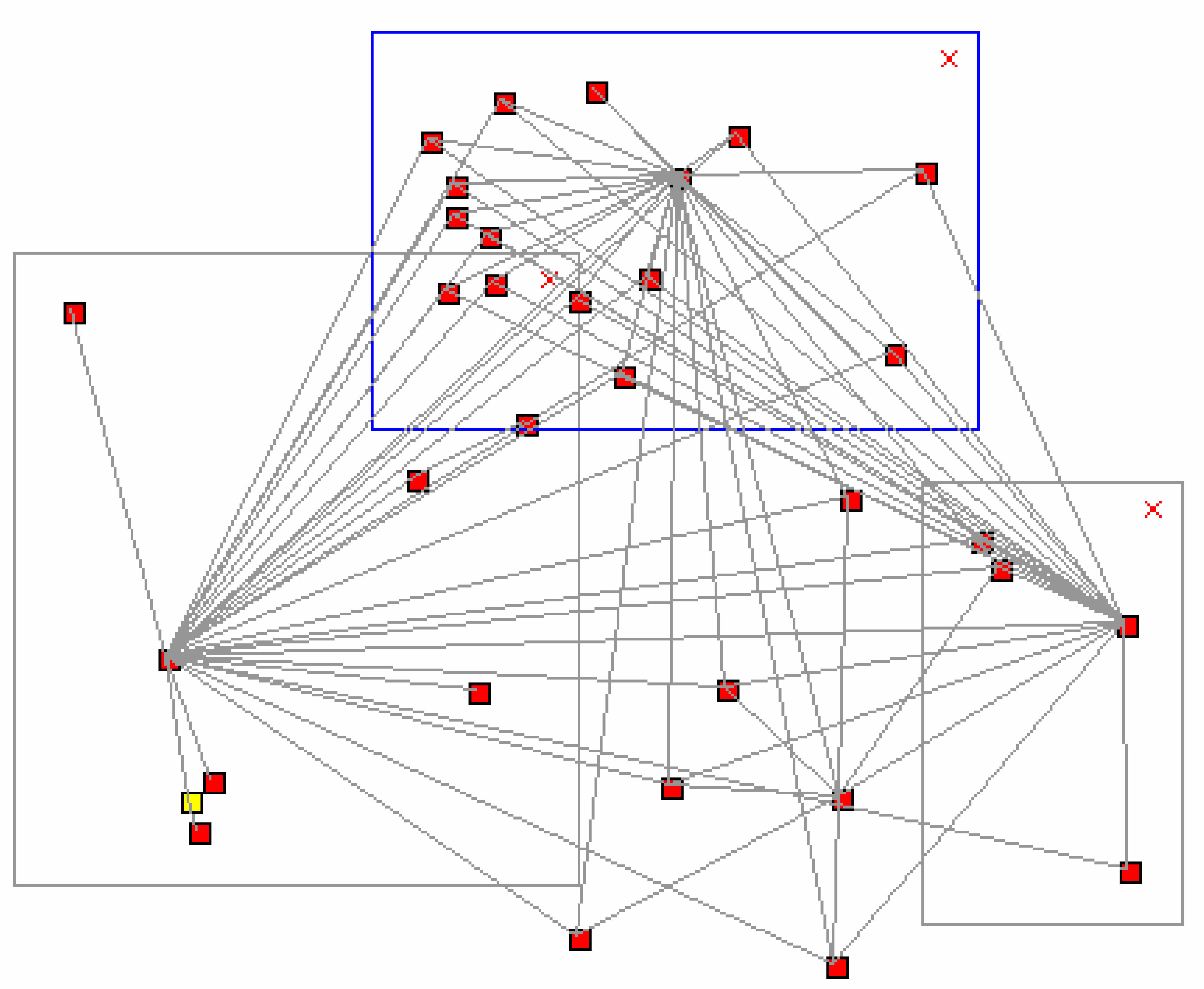
A Course on the Web Graph Chapter 2

The Web Graph

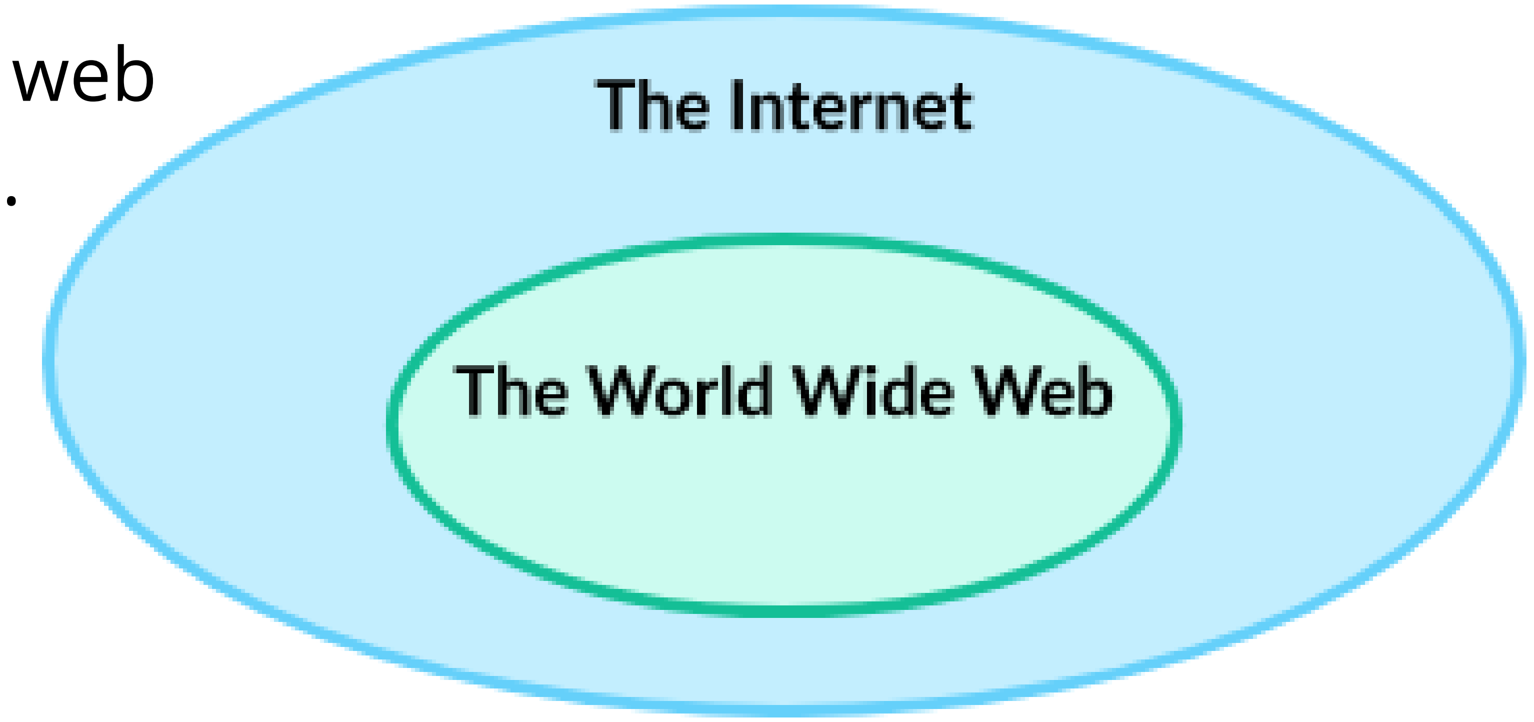


By:
Achyuth Warrier
Ashik Stenny

Basic Definitions

World Wide Web

The World Wide Web (WWW) is a system of interlinked web pages and resources that are accessed via the internet.



Internet:

The internet is the more general term, including various physical components and hardware, as well as other aspects such as web and e-mail.

World Wide Web and The Web Graph

More on World Wide Web

The world wide web consists of information stored and available on the internet.

Most web pages are Hypertext Markup Language (HTML) doc uments identified with strings called Uniform Resource Locators (URLs).

HTML documents are joined by links (or hyperlinks)

The Web Graph

The Web Graph is a mathematical representation of the World Wide Web (WWW) as a directed graph where:

- Webpages = Nodes (Vertices)
- Hyperlinks = Directed Edges (from one webpage to another)

Properties of the Web Graph

- Its order
- Power law degree distribution
- The small world property
- Community structure

How big is the Web

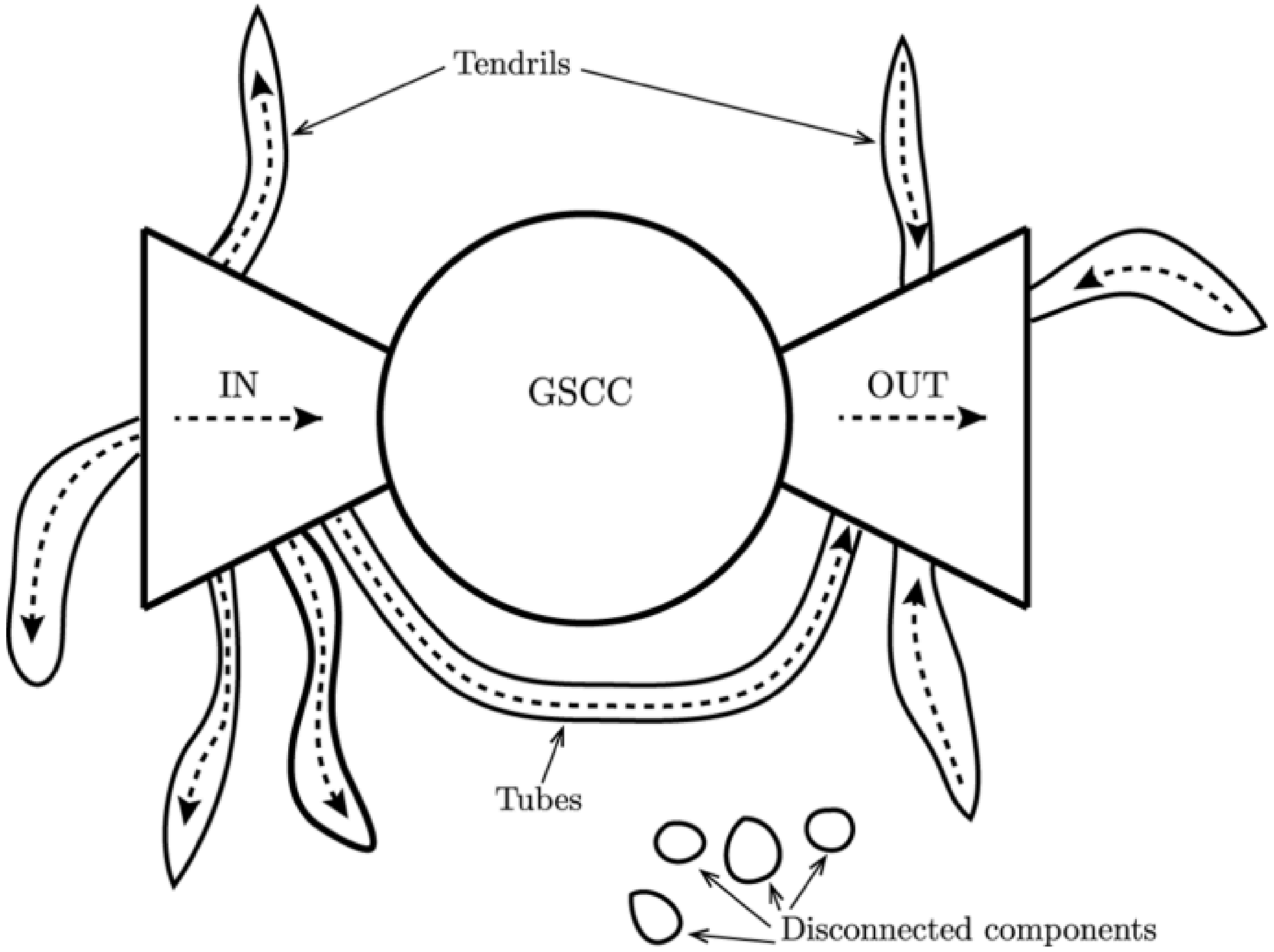
Even a casual surf on the web will convince you that there are an enormous number of web pages and links.

The web has grown rapidly since the mid-1990s. In 1997, it had at least **320 million** pages, which grew to **800 million** by 1999.

By 2005, estimates suggested **11.5 billion** pages, and a later study found **53.7 billion**, with **34.7 billion** indexed by Google.

How big is the Web

The web is often described as having a bow-tie structure, with a strongly connected core (SCC) linking many pages.
 Initially, about one-third of pages were in the core, but more recent estimates suggest over two-thirds, reflecting increased connectivity.



The web graph is considered a sparse graph

Average Degree

$$\frac{1}{|V(G)|}\sum_{v\in V(G)}\deg_G(v)=\frac{2|E(G)|}{|V(G)|}.$$

Based on the average degree graphs can be separated into **sparse** and **dense** graphs

We say that a graph G is sparse if the average degree of G is at most E |V (G)| where E < 1.
Otherwise, we say that G is dense.

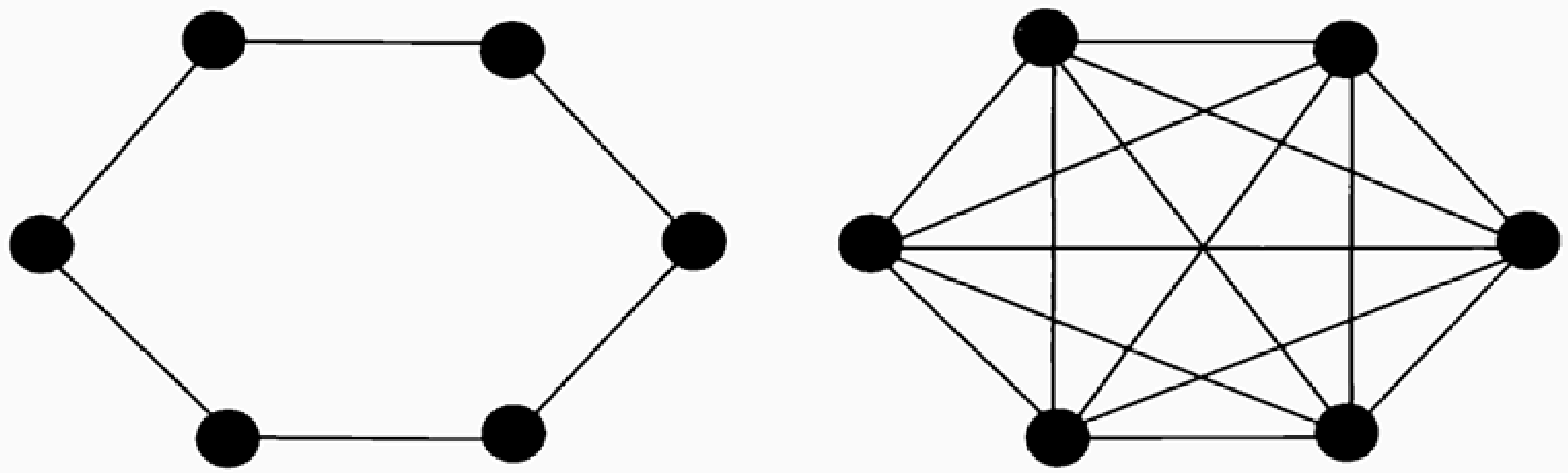


Figure 2.1. A sparse and a dense graph.

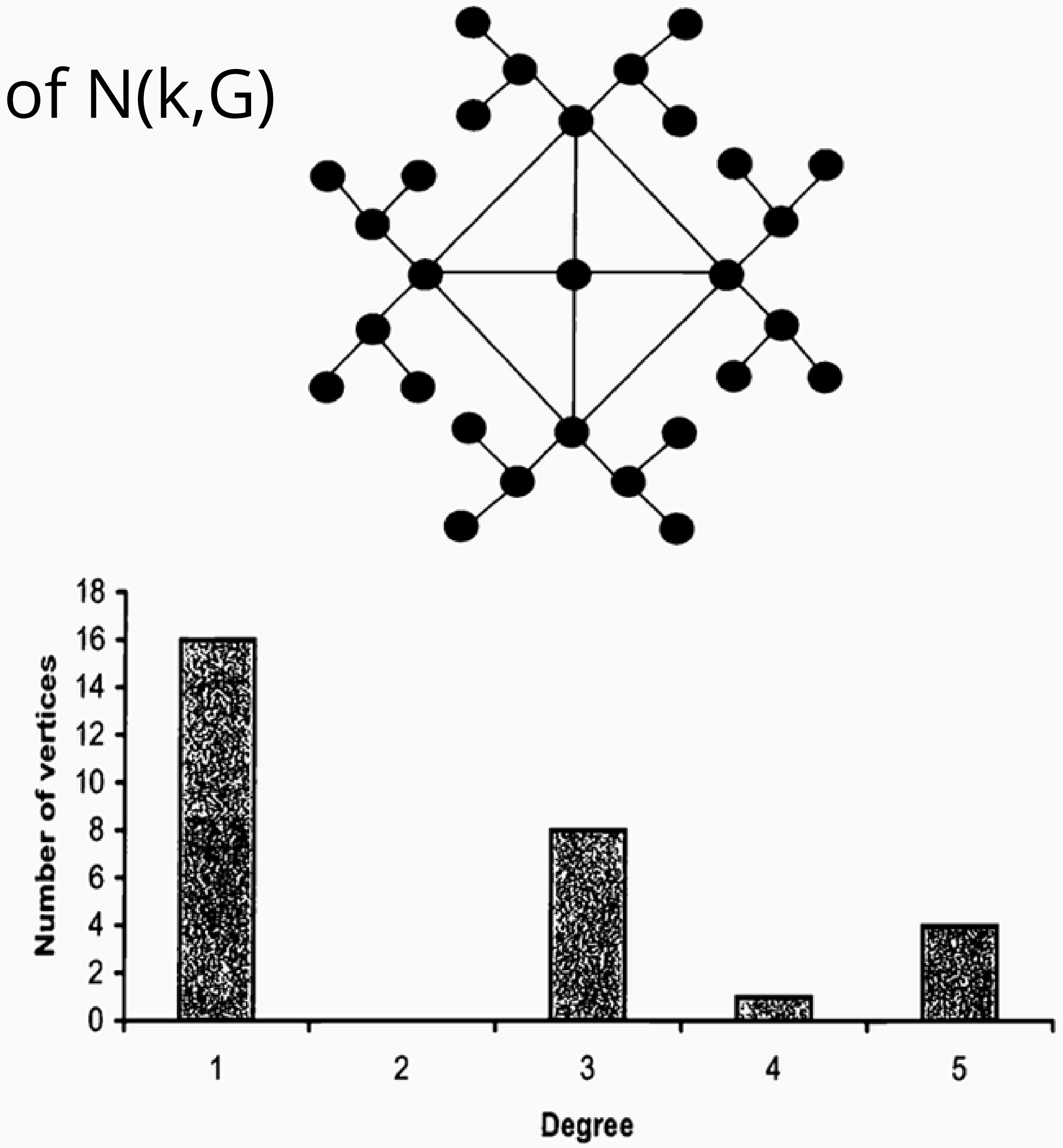
Power - Law Degree Distributions

Given an undirected graph G and a non-negative integer k,

We define N(k,G) by N(k,G) = cardinality of the set of all elements of V with degree equal to k.

For simplicity let us assume the cardinality of V is t then the value of N(k,G) is some integer between from the set {0,1,2....,t}

The parameter Nk,G is the number of vertices of degree k in G



Power - Law Degree Distributions

We say that the degree distribution of G follows a power law
if for each degree k

$$\frac{N_{k,G}}{t} \sim k^{-\beta},$$

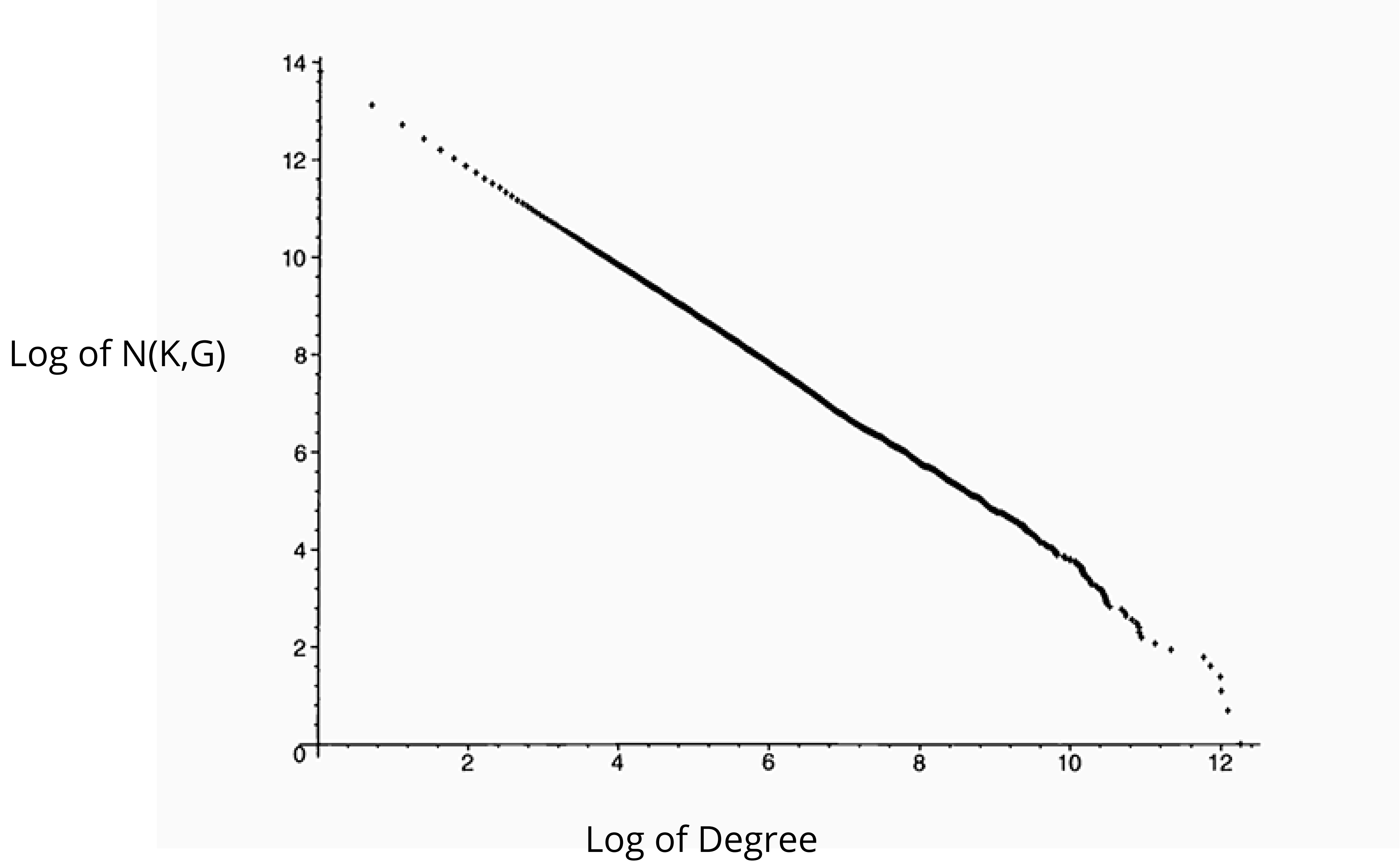
for a fixed real constant β> 1.

The relation is asymptotic because we will be applying this to an infinite family of finite graphs

The reason for this is that since W is a massive graph,
we are more interested in the approximate rather than exact value of N(k,G)

The presence of power law degree distributions reflects a certain undemocratic aspect of W:

While most pages have a small number of links few pages have a large number.



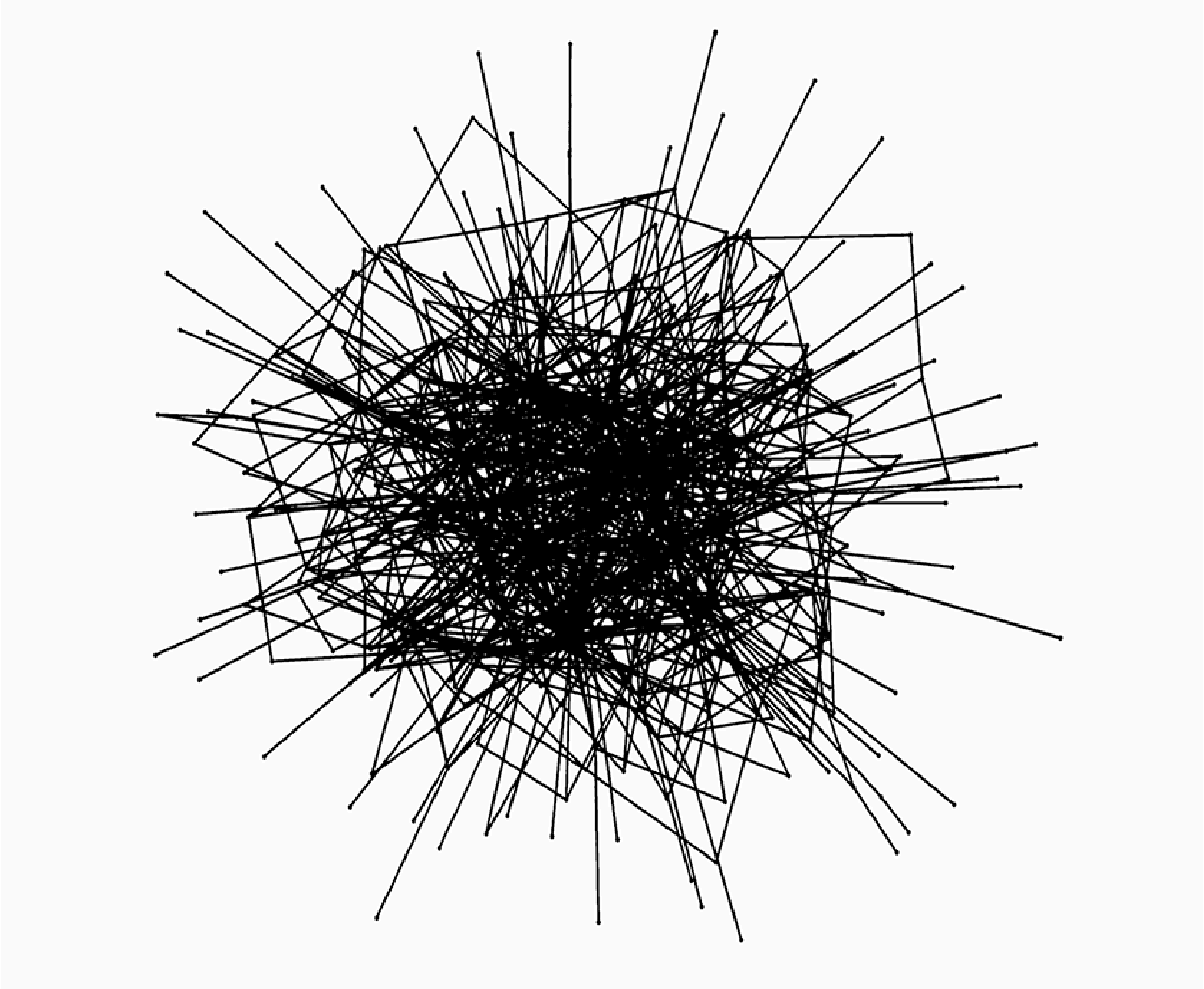


Figure 2.3. A power law graph with 400 vertices.

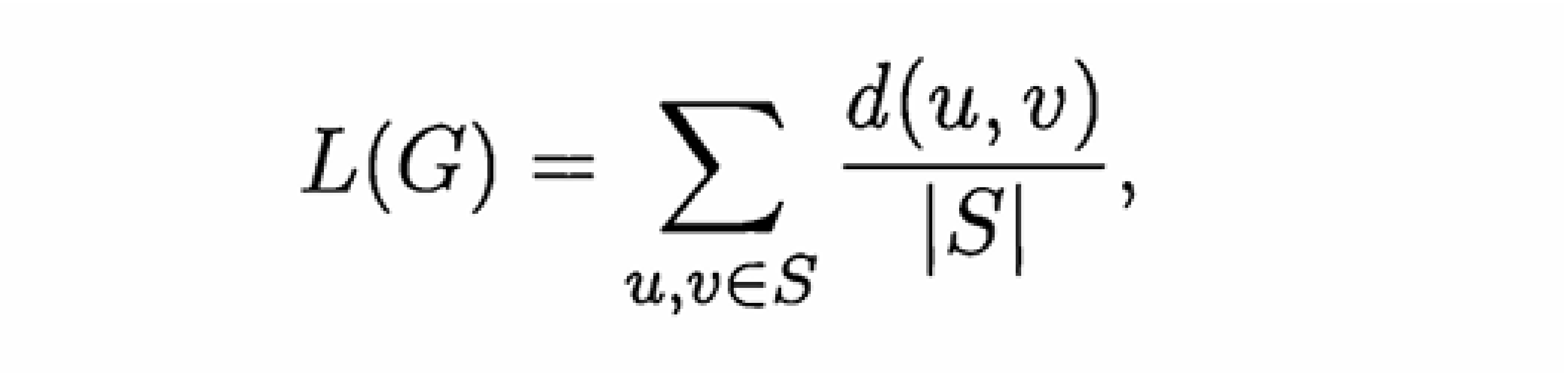
Heavy Tailed Distributions

The small world property

The term small world graphs was first introduced by social scientists Watts and Strogatz [194] in their study of various real-world networks, such as the network of Hollywood movie actors.

The paper [194] introduced the average distance (or characteristic path length) which measures global distances in a graph, and the clustering coefficient, which is a measure of "cliquishness" of neighbourhoods in a graph.

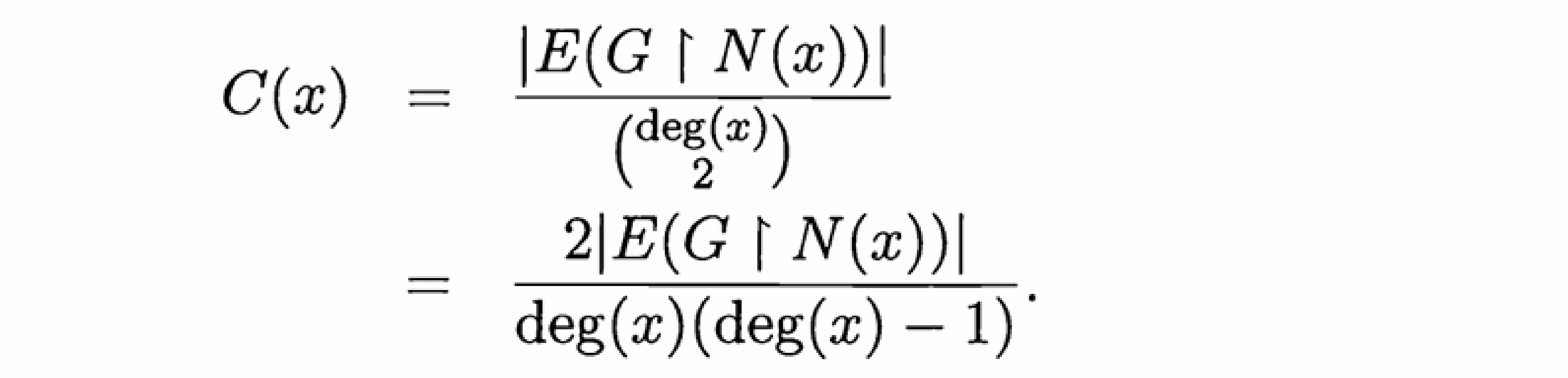
Average Distance



$$L(G) = \sum_{u,v \in S} \frac{d(u,v)}{|S|},$$

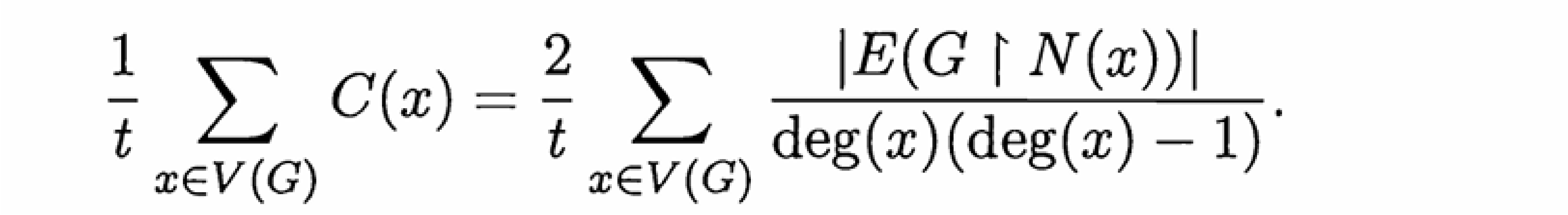
The small world property

For a graph G of order t and x E V (G), define



$$C(x) = \frac{|E(G \restriction N(x))|}{\binom{\deg(x)}{2}} = \frac{2|E(G \restriction N(x))|}{\deg(x)(\deg(x)-1)}.$$

The clustering coefficient of G, written C(G), is the average of the clustering coefficients over all vertices, and so equals



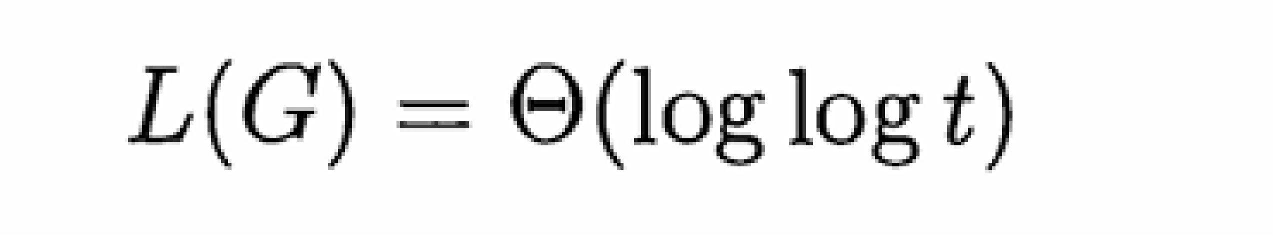
$$\frac{1}{t}\sum_{x \in V(G)} C(x) = \frac{2}{t}\sum_{x \in V(G)} \frac{|E(G \restriction N(x))|}{\deg(x)(\deg(x)-1)}.$$

The small world property

Small world graphs G of order t should satisfy diam(G) = O(log t)

But the observed diam(G) is >900

Better measure of the property from average distances



$$L(G) = \Theta(\log \log t)$$

Community Structure

The web contains many **communities**: sets of pages sharing a common interest or topic. However, there is no consensus for a precise definition of a community in the web graph.

Communities in the web are characterized by dense directed bipartite subgraphs. A **bipartite core** is a directed graph which contains at least one directed bipartite clique as a subgraph, where the directed edges in the subgraph all have tails of one fixed **vertex class**

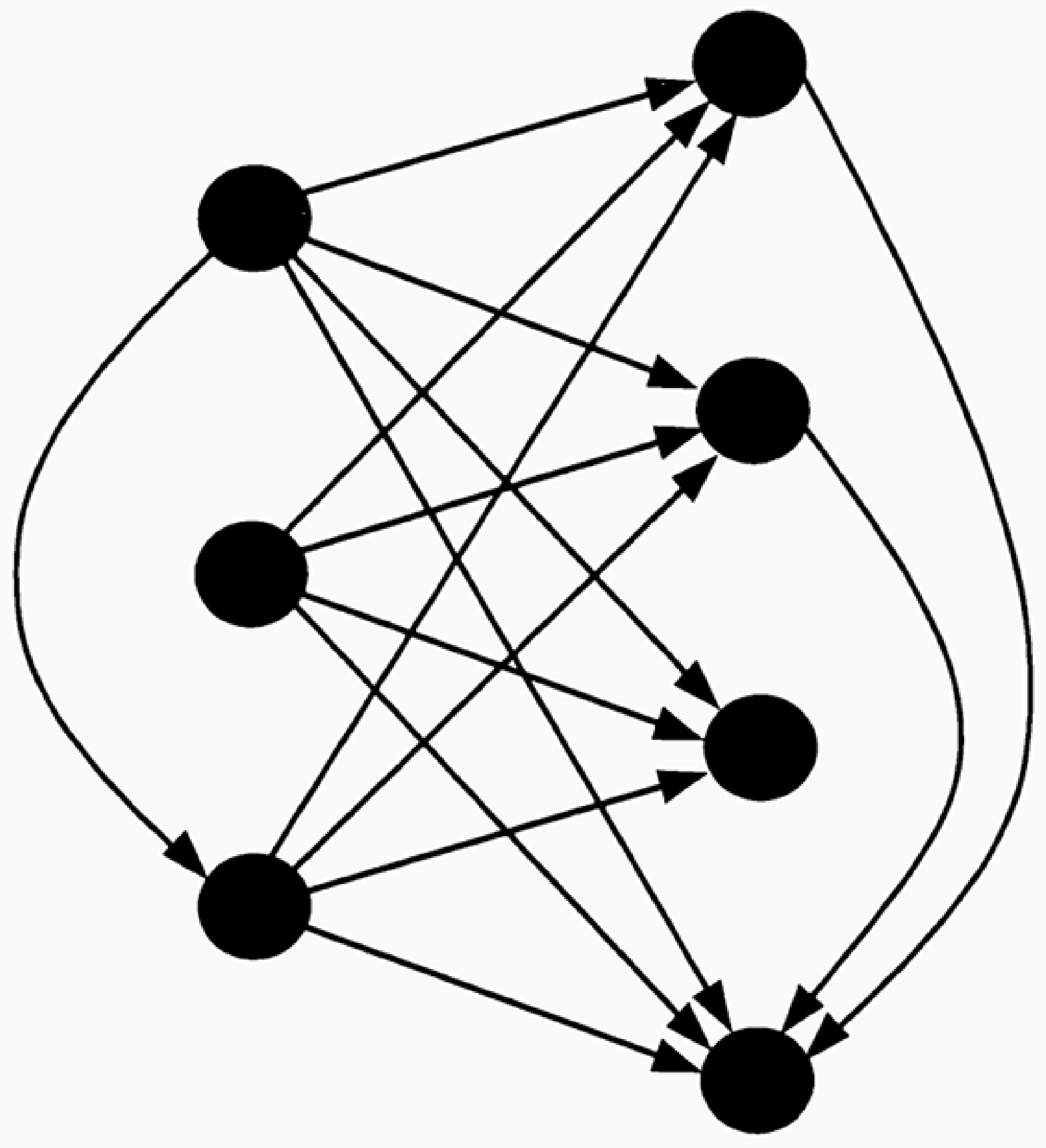


Figure 2.7. A bipartite core.

Community Structure

Why is This Important?

Community Detection: Identifying bipartite cores helps find communities within large networks.

Spam Detection: Many spam structures form artificial dense bipartite cliques (e.g., fake reviews).

Web Structure Analysis: Helps understand how users interact with content on the web.

Other Real-World Self-Organizing Networks

In the late 1990s, research on the properties of W coincided with growing interest in real-world networks that share similar characteristics. These self-organizing networks arise in various fields such as biology, computer science, and social science. They are typically massive, sparse graphs with a power law degree distribution and a small-world structure.

Self-organizing networks are categorized into three types:

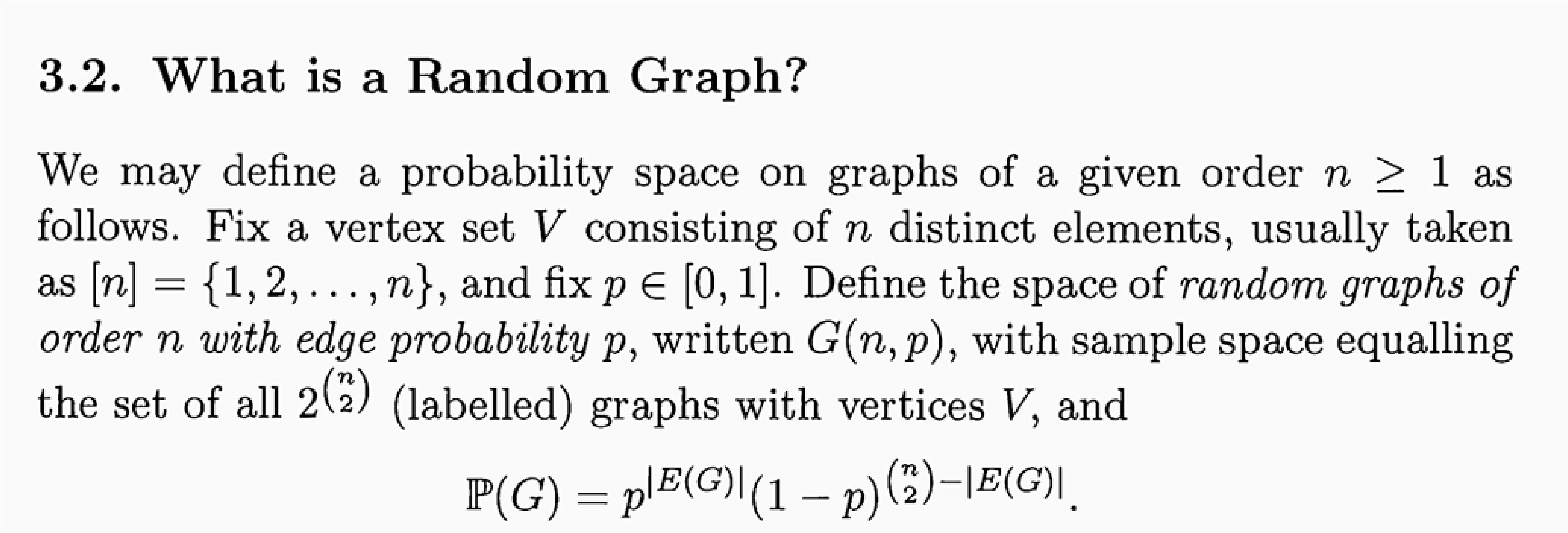
1. Technological
2. Social
3. Biological

The study of W overlaps with broader research on self-organizing networks, but the book primarily focuses on W while providing a brief summary of these other network types.

Social Network

A Course on the Web Graph Chapter 3

Random Graphs



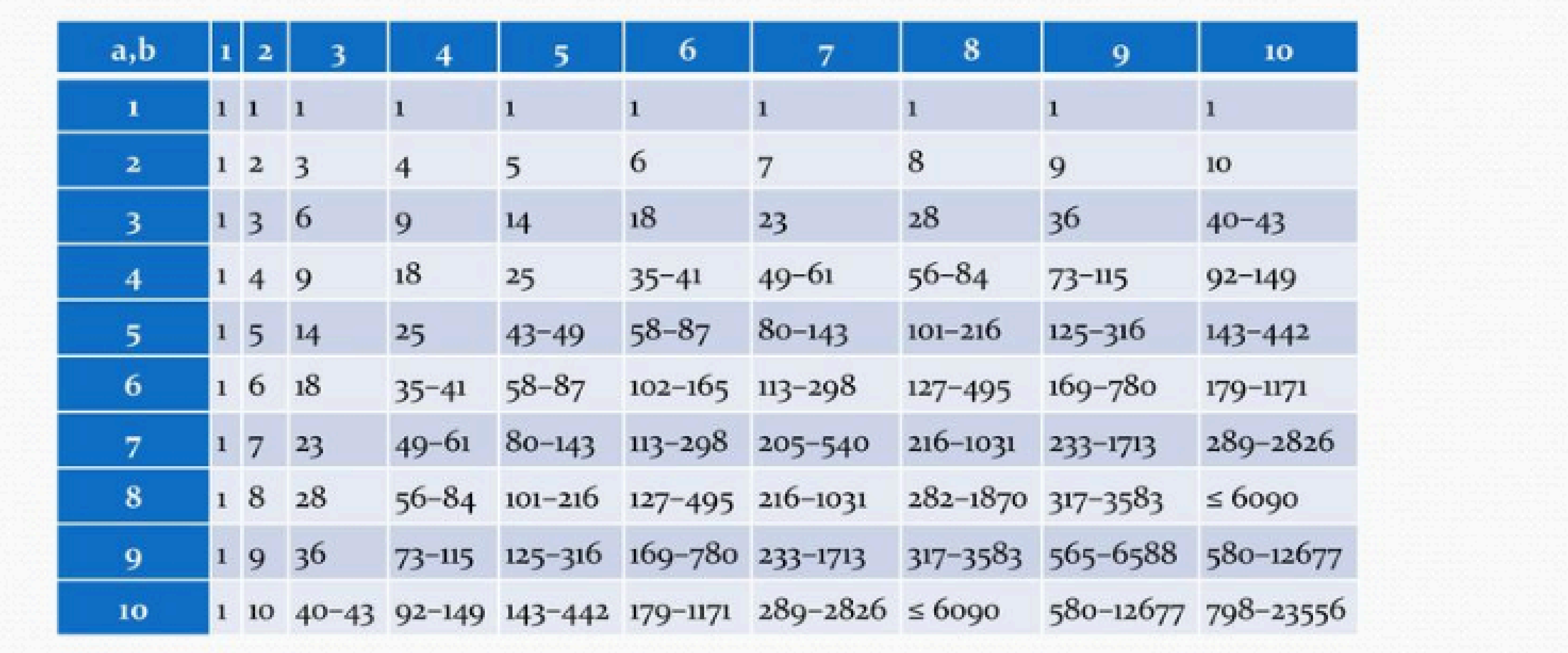
3.2. What is a Random Graph?

We may define a probability space on graphs of a given order $n \geq 1$ as follows. Fix a vertex set V consisting of n distinct elements, usually taken as $[n] = \{1, 2, \ldots, n\}$, and fix $p \in [0, 1]$. Define the space of *random graphs of order n with edge probability p*, written $G(n, p)$, with sample space equalling the set of all $2^{\binom{n}{2}}$ (labelled) graphs with vertices V, and

$$\mathbb{P}(G) = p^{|E(G)|}(1 - p)^{\binom{n}{2} - |E(G)|}.$$

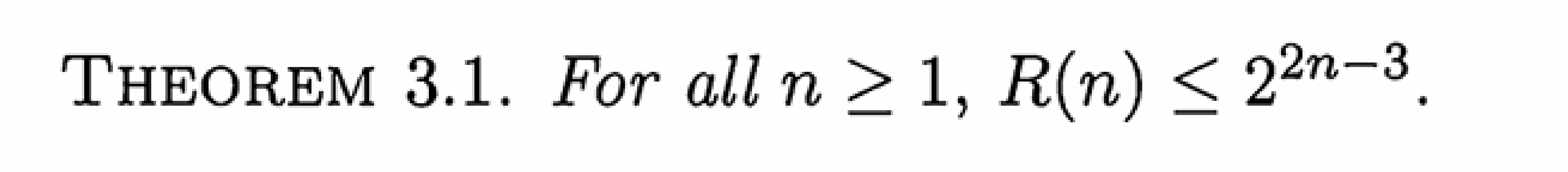
Ramsey Numbers

Ramsey number R(n) is the least integer m such that any graph of order m contains Kn or ~Kn as induced subgraphs.



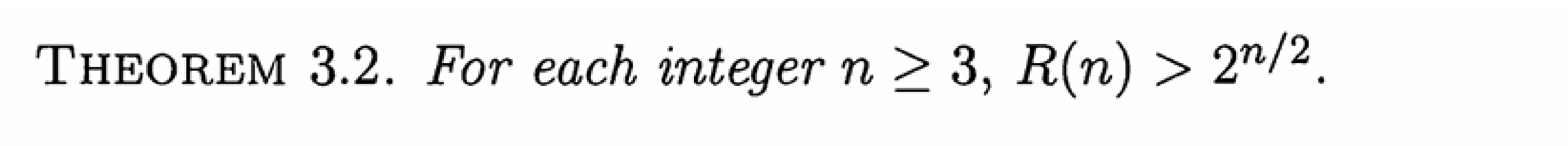
a,b	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	2	3	4	5	6	7	8	9	10
3	1	3	6	9	14	18	23	28	36	40–43
4	1	4	9	18	25	35–41	49–61	56–84	73–115	92–149
5	1	5	14	25	43–49	58–87	80–143	101–216	125–316	143–442
6	1	6	18	35–41	58–87	102–165	113–298	127–495	169–780	179–1171
7	1	7	23	49–61	80–143	113–298	205–540	216–1031	233–1713	289–2826
8	1	8	28	56–84	101–216	127–495	216–1031	282–1870	317–3583	≤ 6090
9	1	9	36	73–115	125–316	169–780	233–1713	317–3583	565–6588	580–12677
10	1	10	40–43	92–149	143–442	179–1171	289–2826	≤ 6090	580–12677	798–23556

Existence of Ramsey Numbers



THEOREM 3.1. *For all* $n \geq 1$, $R(n) \leq 2^{2n-3}$.

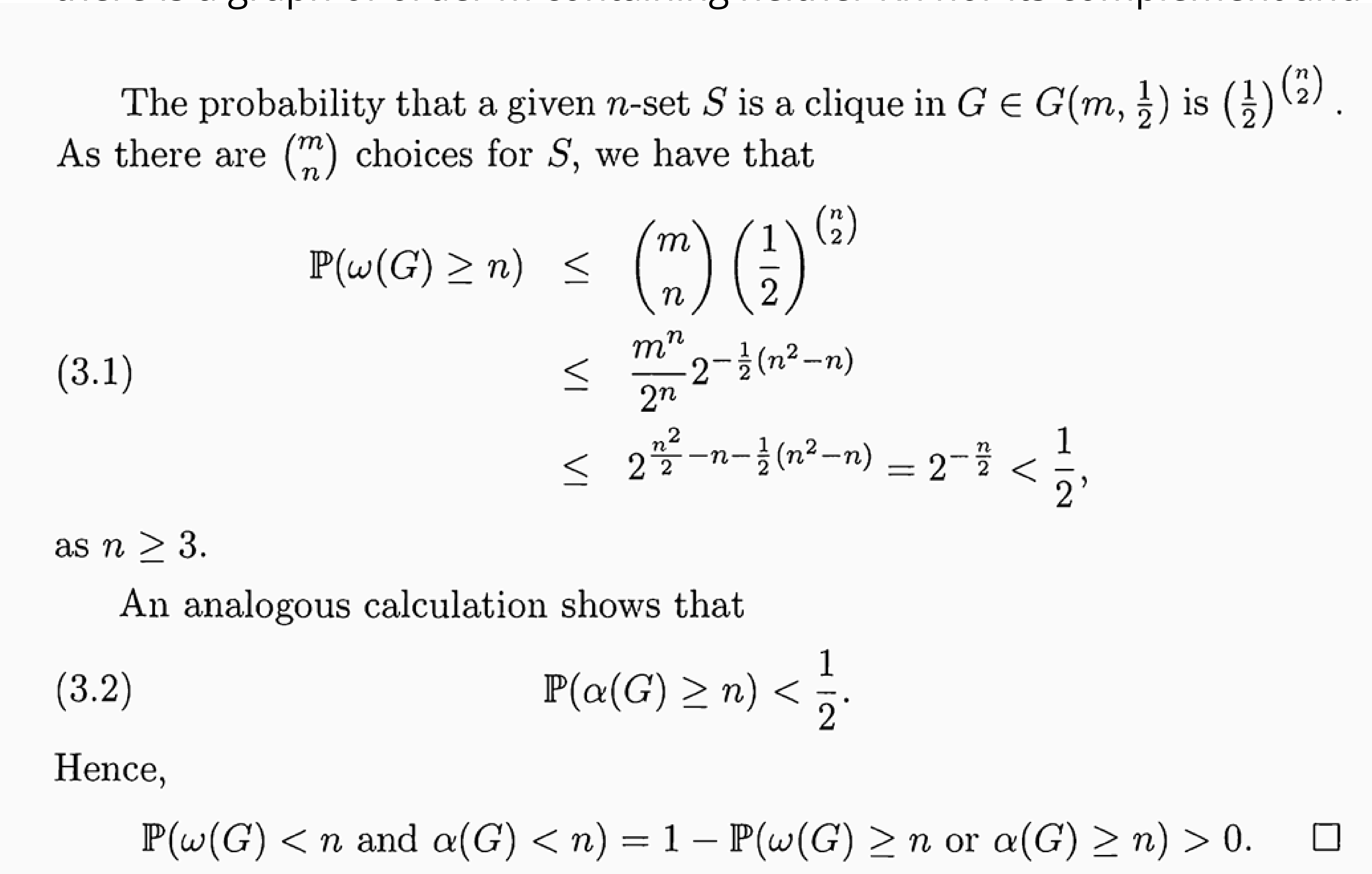
Application of random graphs



THEOREM 3.2. *For each integer* $n \geq 3$, $R(n) > 2^{n/2}$.

Proof:

We prove that for a fixed m < 2^(n/2), with positive probability,
G element of G(m, 1/2) satisfies alpha(G) < n and w(G) < n. Hence,
there is a graph of order m containing neither Kn nor its complement and the proof follows.



The probability that a given n-set S is a clique in $G \in G(m, \frac{1}{2})$ is $\left(\frac{1}{2}\right)^{\binom{n}{2}}$. As there are $\binom{m}{n}$ choices for S, we have that

$$\begin{aligned} \mathbb{P}(\omega(G) \geq n) &\leq \binom{m}{n}\left(\frac{1}{2}\right)^{\binom{n}{2}} \\ &\leq \frac{m^n}{2^n} 2^{-\frac{1}{2}(n^2-n)} \qquad (3.1) \\ &\leq 2^{\frac{n^2}{2}-n-\frac{1}{2}(n^2-n)} = 2^{-\frac{n}{2}} < \frac{1}{2}, \end{aligned}$$

as $n \geq 3$.

An analogous calculation shows that

$$\mathbb{P}(\alpha(G) \geq n) < \frac{1}{2}. \qquad (3.2)$$

Hence,

$$\mathbb{P}(\omega(G) < n \text{ and } \alpha(G) < n) = 1 - \mathbb{P}(\omega(G) \geq n \text{ or } \alpha(G) \geq n) > 0. \qquad \square$$

Graph Property:

A property preserved by isomorphism;
for example, being planar and possessing Hamilton cycles

A.A.S: Asymptotically almost surely

We say that G element of G(n, p) satisfies Property, P asymptotically almost surely (or a. a. s. for short) if when n tends to infinity Probability(G element of G(n, p) satisfies P) tends to 1.

In this case we say any random G element of G(n, p) satisfies P with high probability (w.h.p)

Adjacency Properties:

Let us look into a few properties which are asymptotically satisfied.

An adjacency property is a global property of a graph asserting that
for every set S of vertices of some fixed type,
there is a vertex joined to some of the vertices of S in a prescribed way

n-e.c Adjacency Property

A graph is n-existentially closed or n-e.c., if
for all disjoint sets of vertices A and B with |A U B| = n (one of A or B can be empty),
there is a vertex z not in A U B joined to each vertex of A and no vertex of B.

Cartesian Product of Graphs

The Cartesian product of two graphs G=(V_G,E_G) and H=(V_H,E_H) denoted as G□HG, is a graph whose vertex set and edge set are defined as follows:

- Vertex Set:
- V(G□H)=V_G×V_H
- This means each vertex in G□H is a pair (g,h) where g∈V_G and h∈V_H.
- Edge Set:
- Two vertices (g_1, h_1) and (g_2, h_2) in G□H are adjacent if and only if:
  - g_1 = g_2 and (h_1,h_2)∈E_H (i.e., the second coordinate changes according to an edge in H), or
  - h_1=h_2 and (g_1,g_2)∈E_G (i.e., the first coordinate changes according to an edge in G).

Properties of Cartesian Product

1. Commutative: G□H≅H□G
2. Associative: (G□H)□K≅G□(H□K)
3. Degree Property: deg(G□H)(v,w)=deg(G)(v)+deg(H)(w)
4. Connectedness: If both G and H are connected, so is G□H.

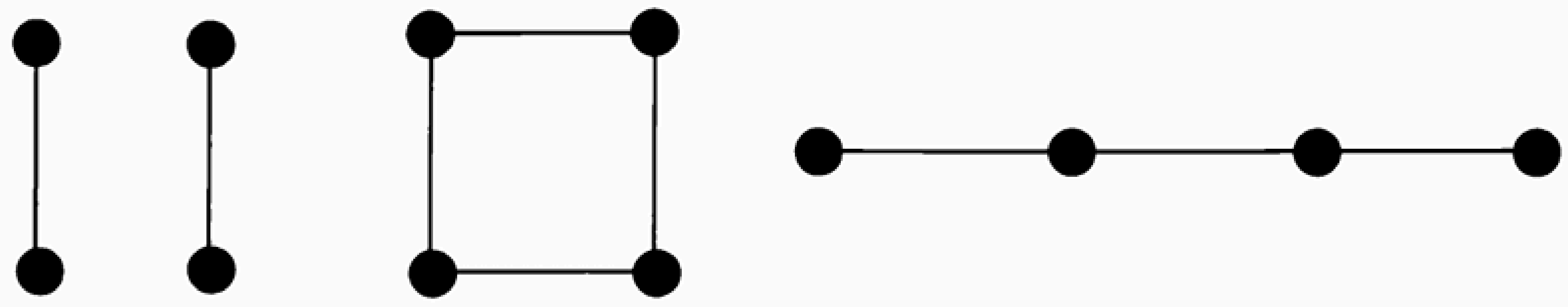


Figure 3.2. The 1-e.c. graphs of order 4.

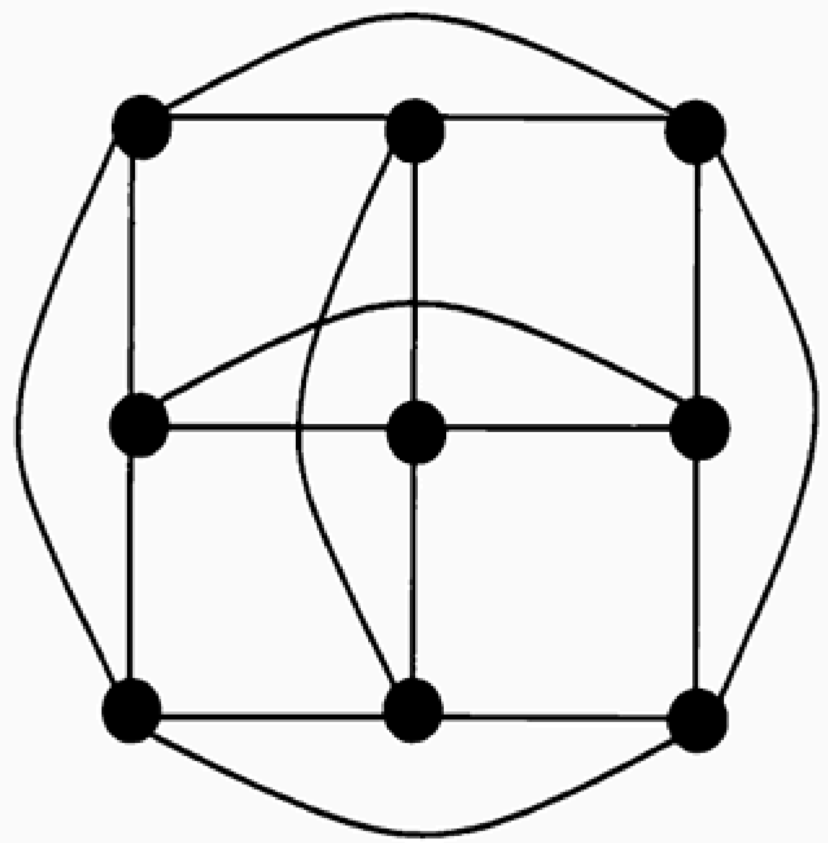
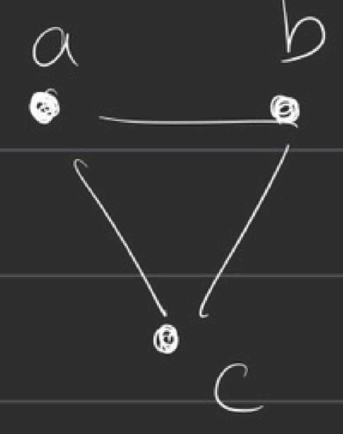
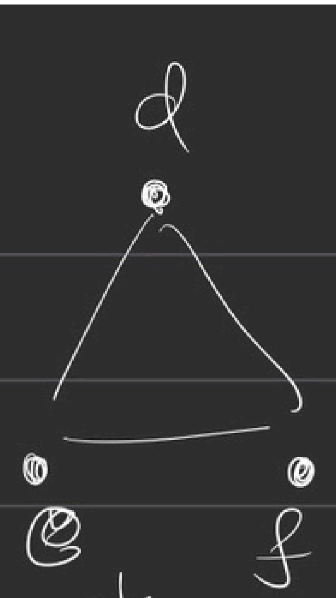


Figure 3.3. The graph $K_3 \square K_3$.

$$20 \leq m_{ec}(3) \leq 28.$$



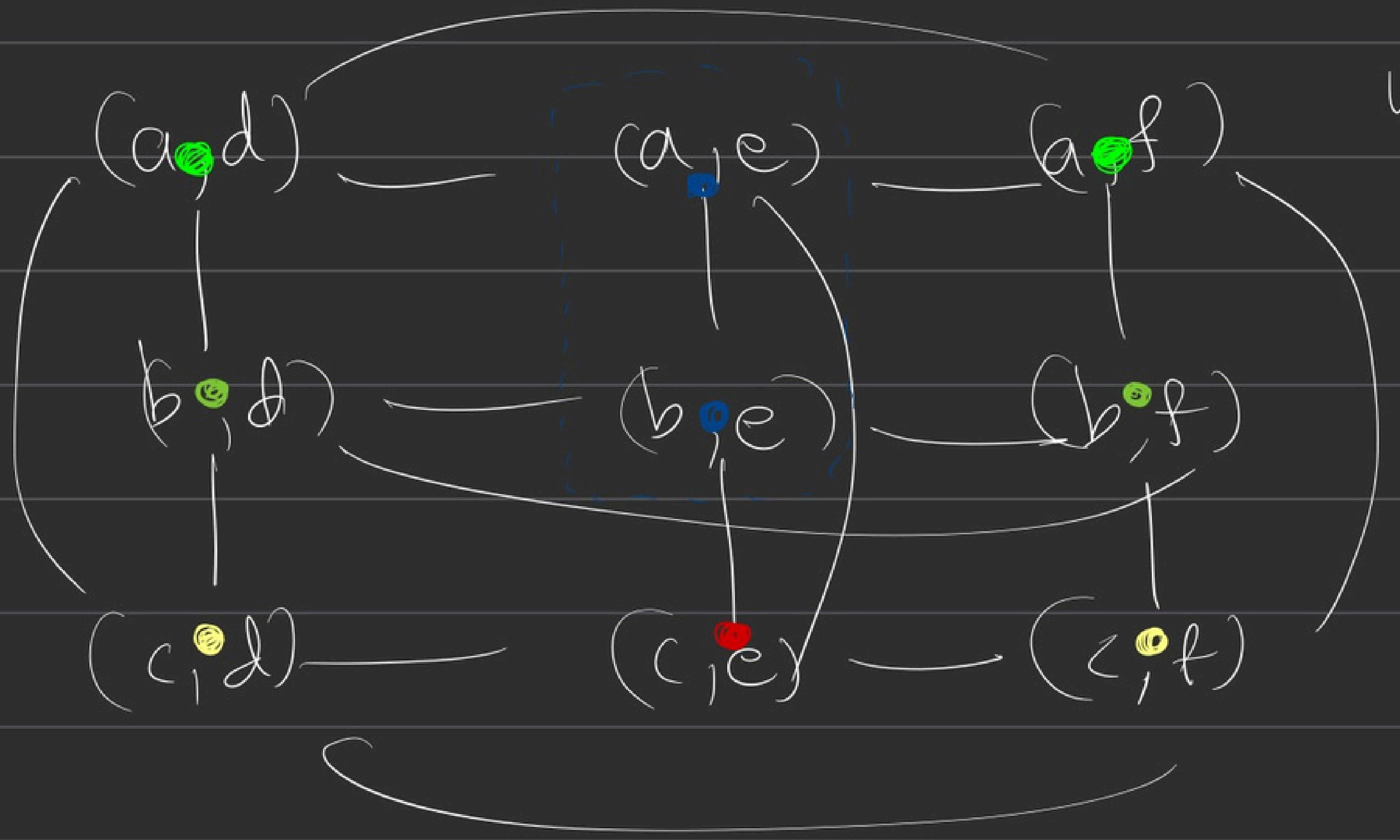
K_3



K_3

Cartesian product $K_3 \square K_3$

$m_{ec}(2) = 9$



we need $2^2 = 4$ vertices for each 2-set

Pseudo Randomness

Pseudo-randomness refers to a graph exhibiting properties
that are characteristic of a truly random graph, even though
it's not constructed randomly but by a deterministic process or algorithm

Quasi Randomness

A graph is considered "quasi-random" if it exhibits many of the
same properties as a truly random graph of the same size and
edge density, even though it might not be generated by a truly random process.

Difference ?

The definition of Quasi randomness is more rigid

Notations involved-Quasi Randomness

To each graph G of order n we may associate its n x n adjacency matrix A(G).

As A(G) is a real-symmetric matrix, it has n real eigenvalues

which are ordered by absolute value: $|\lambda 1|>=|\lambda 2|\ldots >=|\lambda n|$

Ns (H, G) be the number of labelled subgraphs of G isomorphic to H,

while Nis(H, G) is the number of labelled induced subgraphs of G isomorphic to H

For X C V_G, let e(X) is the number of edges in the subgraph induced by X on G.

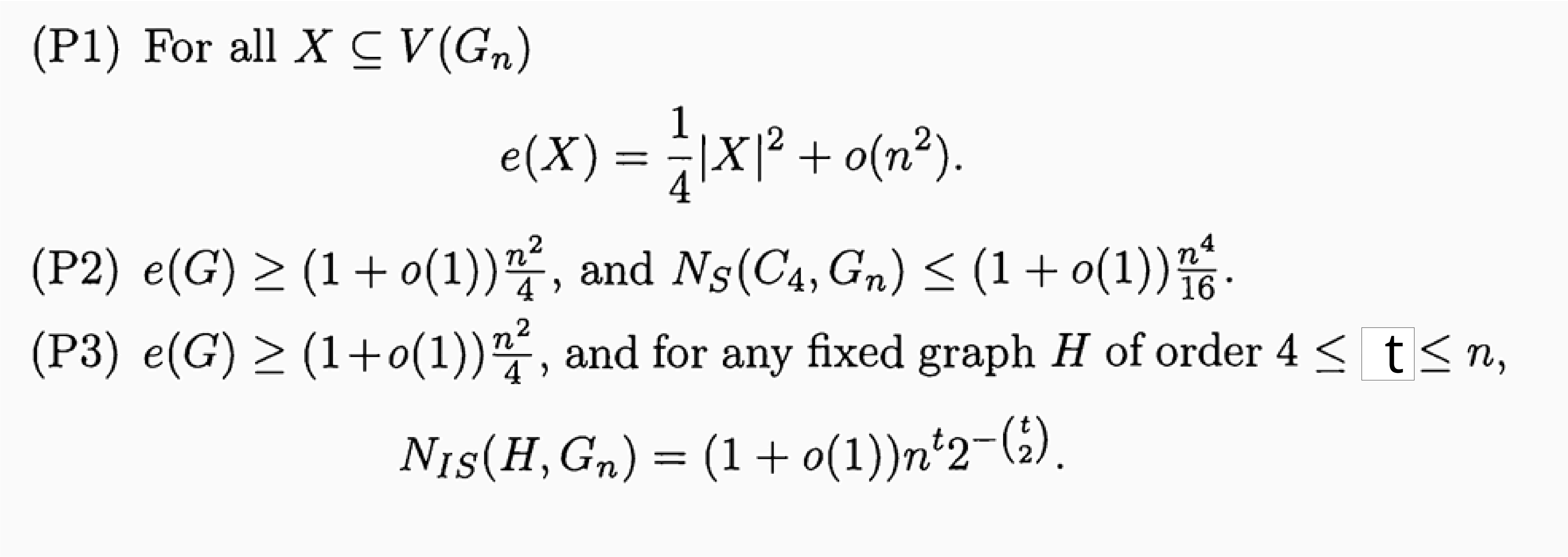
For vertices u and v, define s(u, v) to be the set of vertices joined to both u and v or neither u nor v.

Quasi Randomness - Properties

Any graph that is quasi random satisfies a few properties

These properties all hold a.a.s. in G(n, 1/2)

Any deterministic graph that satisfies any of these properties satisfy all of them



(P1) For all $X \subseteq V(G_n)$

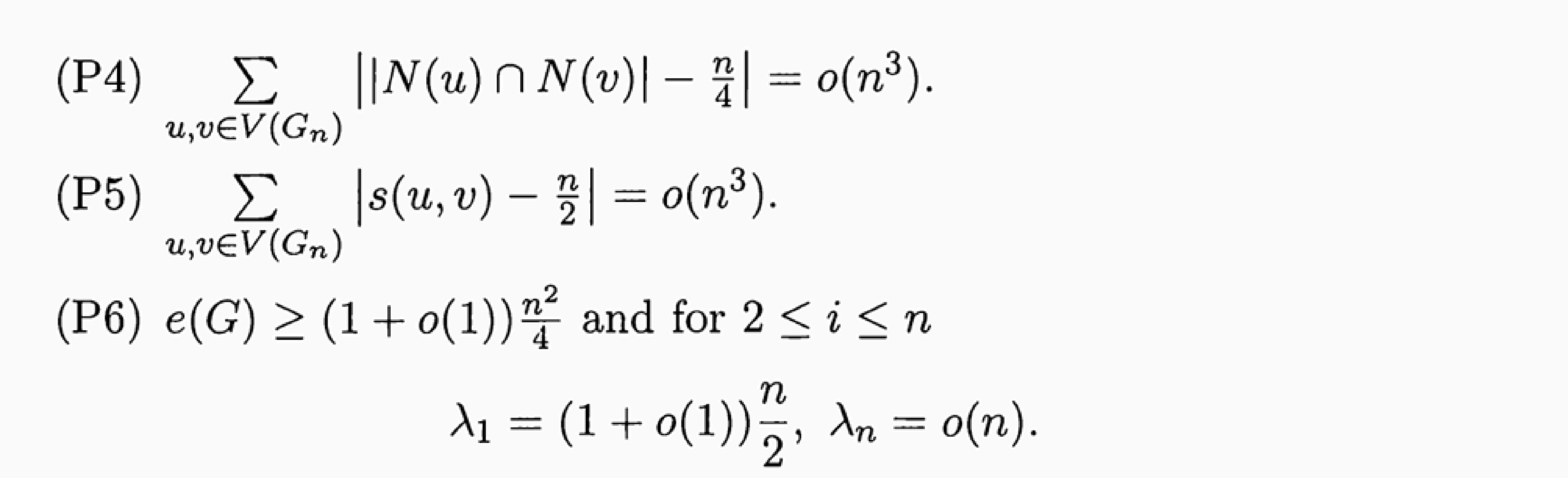
$$e(X) = \frac{1}{4}|X|^2 + o(n^2).$$

(P2) $e(G) \geq (1+o(1))\frac{n^2}{4}$, and $N_S(C_4, G_n) \leq (1+o(1))\frac{n^4}{16}$.

(P3) $e(G) \geq (1+o(1))\frac{n^2}{4}$, and for any fixed graph H of order $4 \leq t \leq n$,

$$N_{IS}(H, G_n) = (1+o(1))n^t 2^{-\binom{t}{2}}.$$

Quasi Randomness - Properties

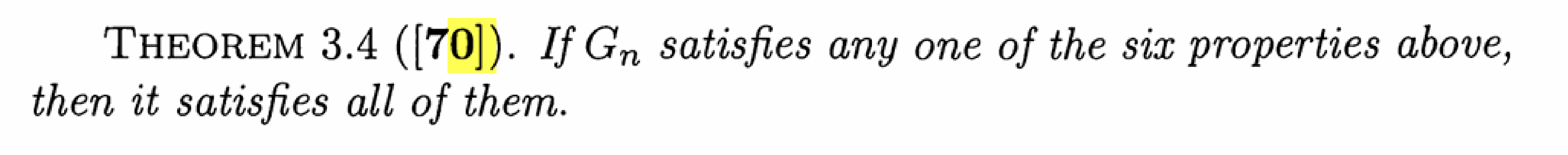


(P4) $$\sum_{u,v \in V(G_n)} \left| |N(u) \cap N(v)| - \tfrac{n}{4} \right| = o(n^3).$$

(P5) $$\sum_{u,v \in V(G_n)} \left| s(u,v) - \tfrac{n}{2} \right| = o(n^3).$$

(P6) $e(G) \geq (1 + o(1))\frac{n^2}{4}$ and for $2 \leq i \leq n$

$$\lambda_1 = (1 + o(1))\frac{n}{2}, \ \lambda_n = o(n).$$

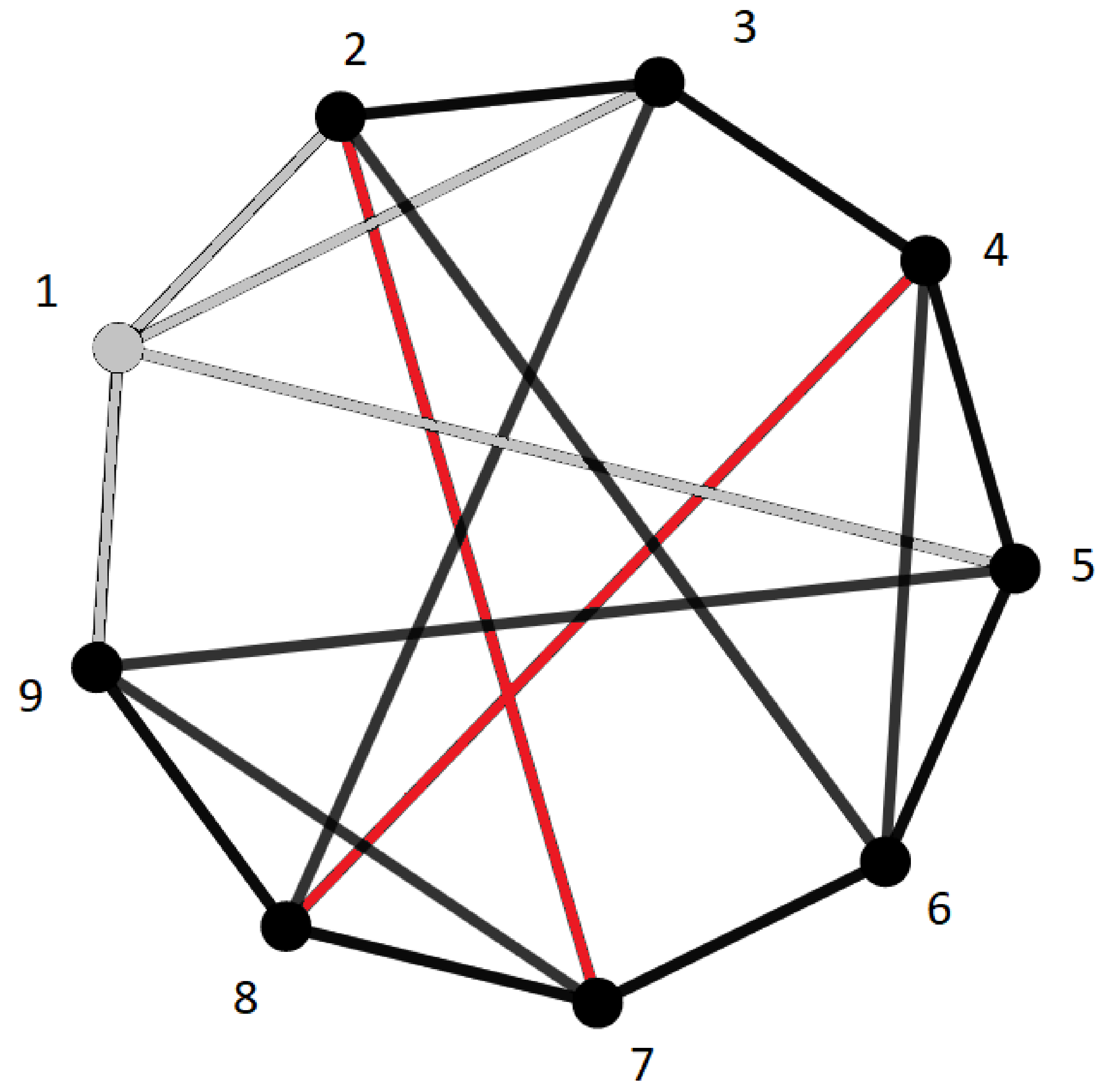


THEOREM 3.4 ([70]). *If G_n satisfies any one of the six properties above, then it satisfies all of them.*

Strongly Regular Graphs(SRG)

A k-regular graph G with v vertices, such that each pair of joined vertices has exactly A common neighbors, and each pair of non-joined vertices has exactly p common neighbors is called a strongly regular graph; we say that G is an SRG (v, k, A, p)

Example - Paley graphs



Paley Graphs

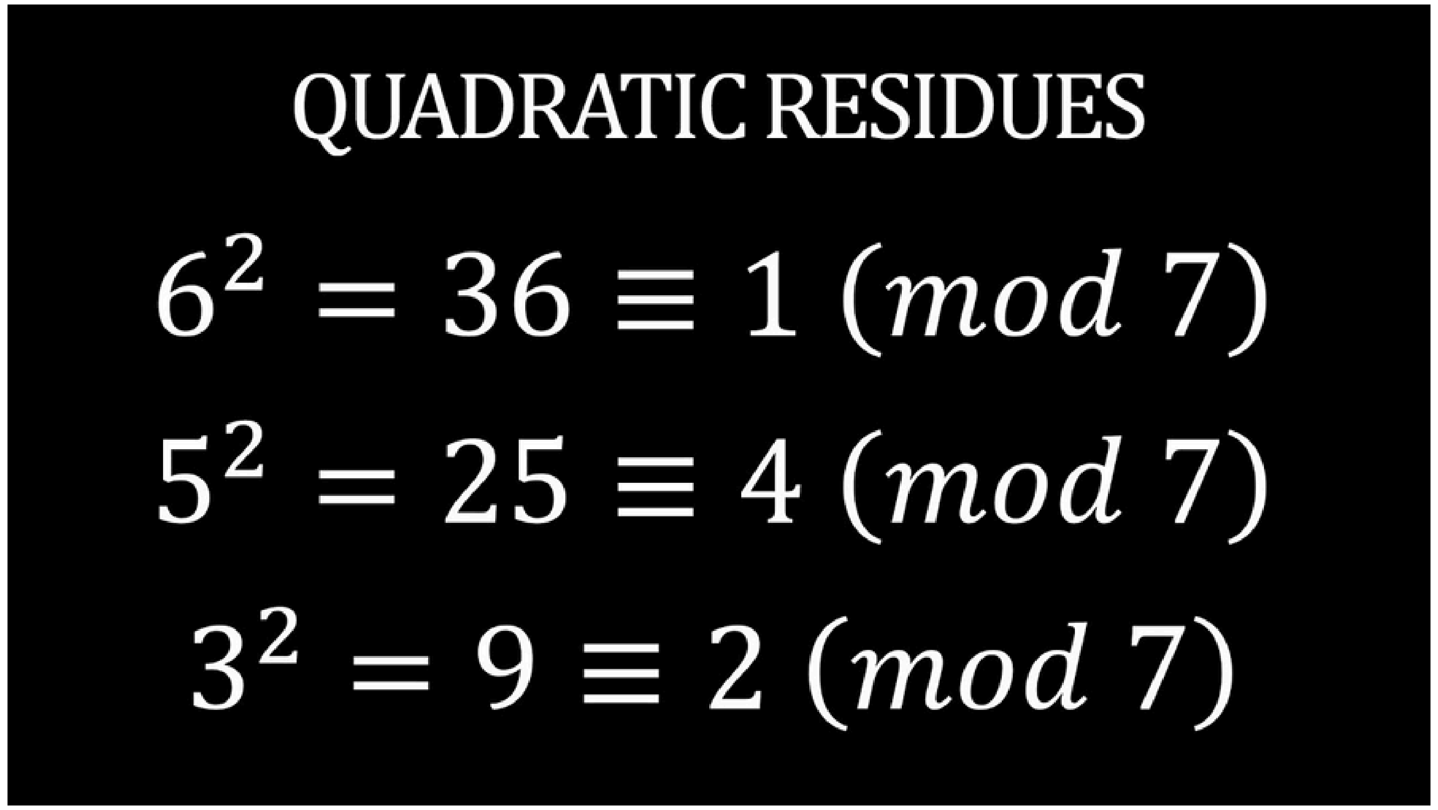
Quadratic Residue

A quadratic residue modulo an integer n (typically a prime p) is an integer a such that there exists a non-zero integer x satisfying

$x^2 \equiv a(modn)$.

In other words, a is a quadratic residue modulo n if it is congruent to a perfect square modulo n. If no such x exists, then a is called a quadratic non-residue.

For example consider n = 7. The set of quadratic residues is {1,2,4}, while the set of quadratic non-residues is {3,5,6}



QUADRATIC RESIDUES

$$6^2 = 36 \equiv 1\ (mod\ 7)$$

$$5^2 = 25 \equiv 4\ (mod\ 7)$$

$$3^2 = 9 \equiv 2\ (mod\ 7)$$

Paley Graphs

Vertex Set:

Choose a positive integer q (often a prime or a prime power) such that

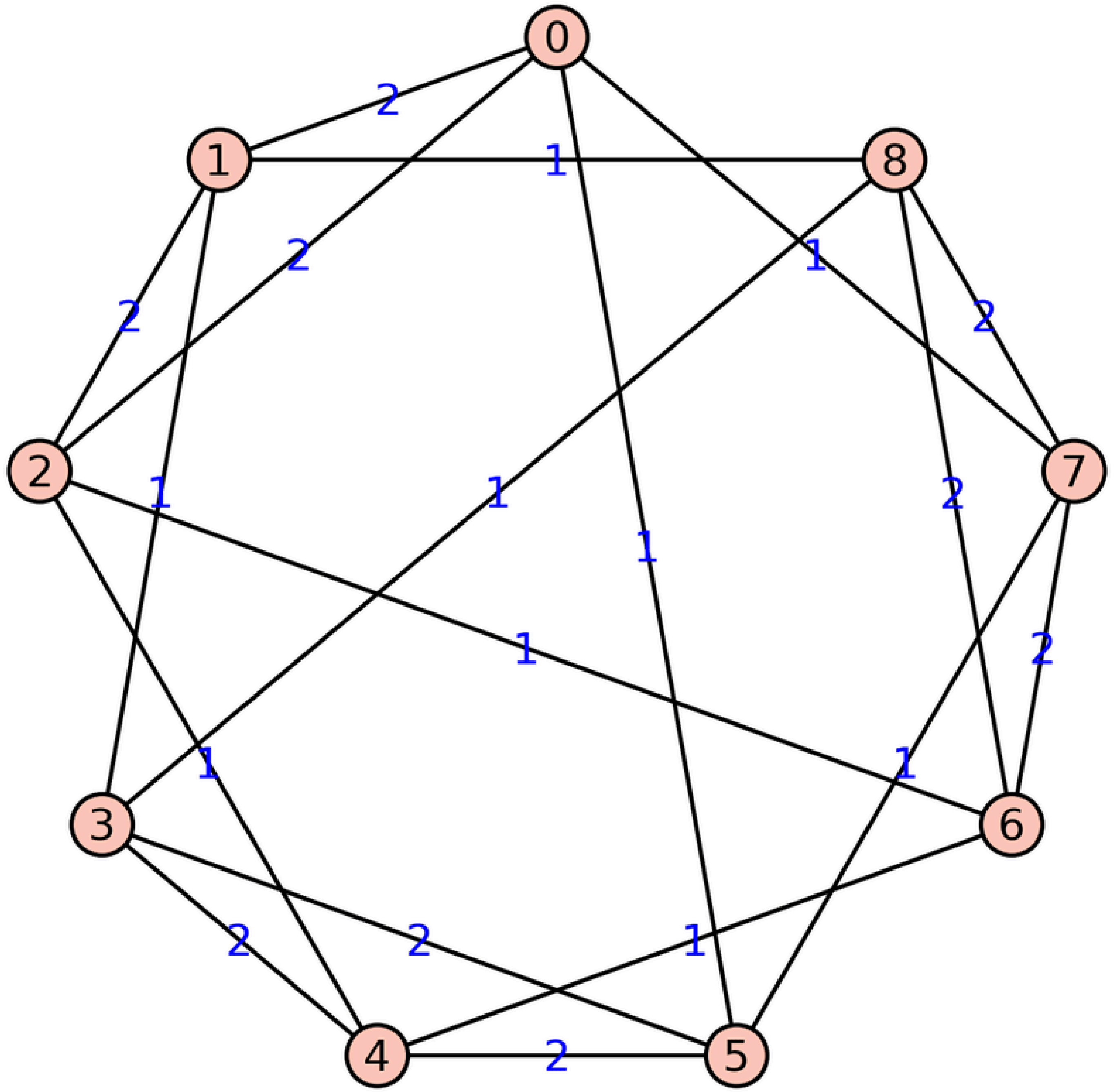
$q \equiv 1 \pmod{4}$.

Label the vertices with the numbers

$0,1,2,\ldots,q-1$.

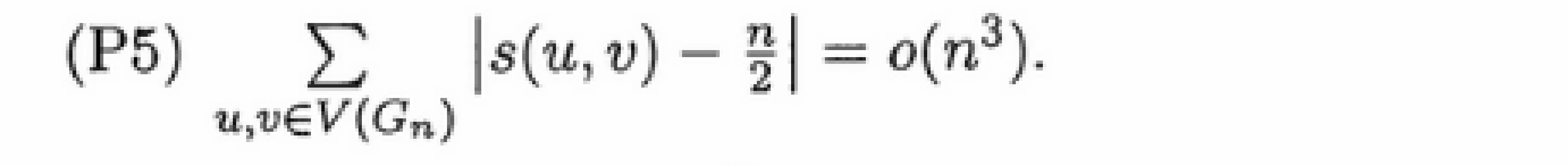
Edge Set:

For any two distinct vertices i and j (considered modulo q), connect i and j with an edge if and only if the difference $(i-j)$ (computed modulo q) is a quadratic residue modulo q.



Paley Graph - Quasi Random

Paley Graphs are Quasi Random because they satisfy P5



$$\text{(P5)} \quad \sum_{u,v \in V(G_n)} \left| s(u,v) - \tfrac{n}{2} \right| = o(n^3).$$

Proof:

Let us consider an element z of the field for it to be vertex contributing to s(u,v) it has to be satisfy:
either
1)z-u is a quadratic residue and z-v is a quadratic residue
or
2)z-u is not a quadratic residue and z-v is not a quadratic residue

Paley Graph - Quasi Random

Proof:

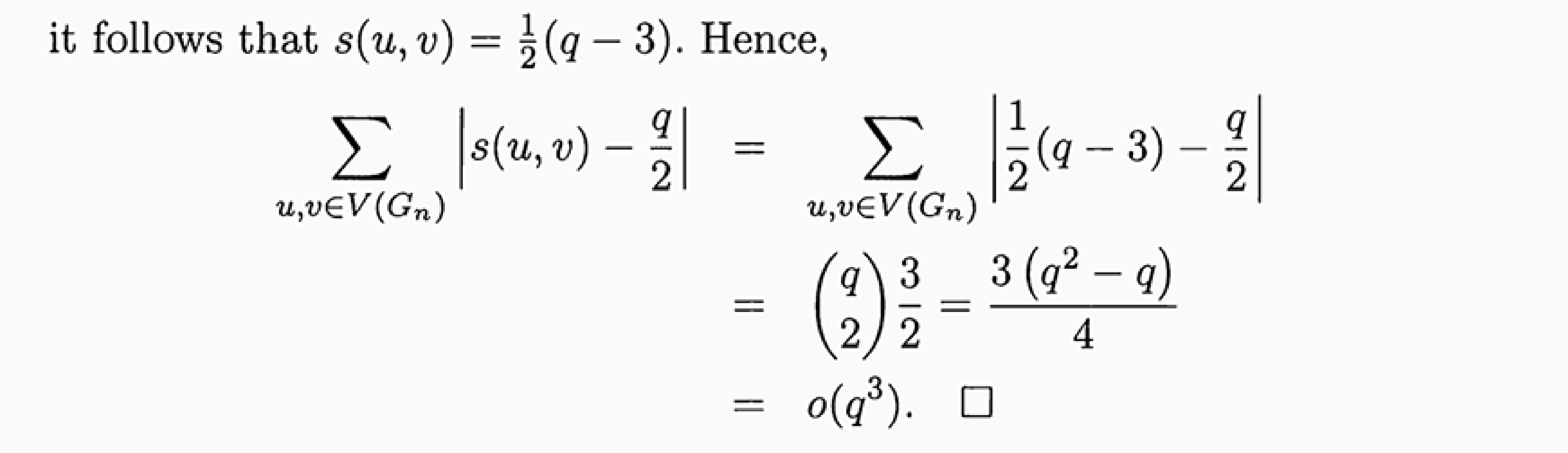
Let us consider an element z of the field for it to be vertex contributing to s(u,v) it has to be satisfy:
either
1)z-u is a quadratic residue and z-v is a quadratic residue
or
2)z-u is not a quadratic residue and z-v is not a quadratic residue
We can include both these conditions in one expression, following term must be a quadratic residue

$$\frac{z-u}{z-v} = a \qquad\qquad \frac{z-u}{z-v} = 1 + \frac{v-u}{z-v} = a,$$

Here there exists a 1 to 1 relation between a and z and a is a square
No of squares in a the field is (q-1)/2 and we need to subtract 1 because a can not take
value 1

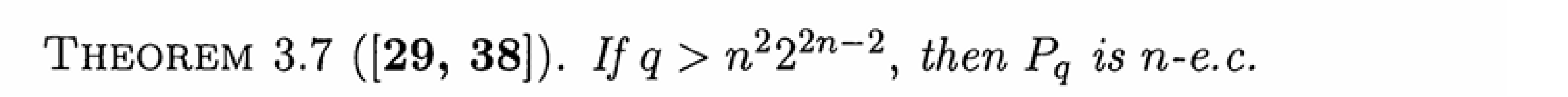
Paley Graph - Quasi Random

Proof:

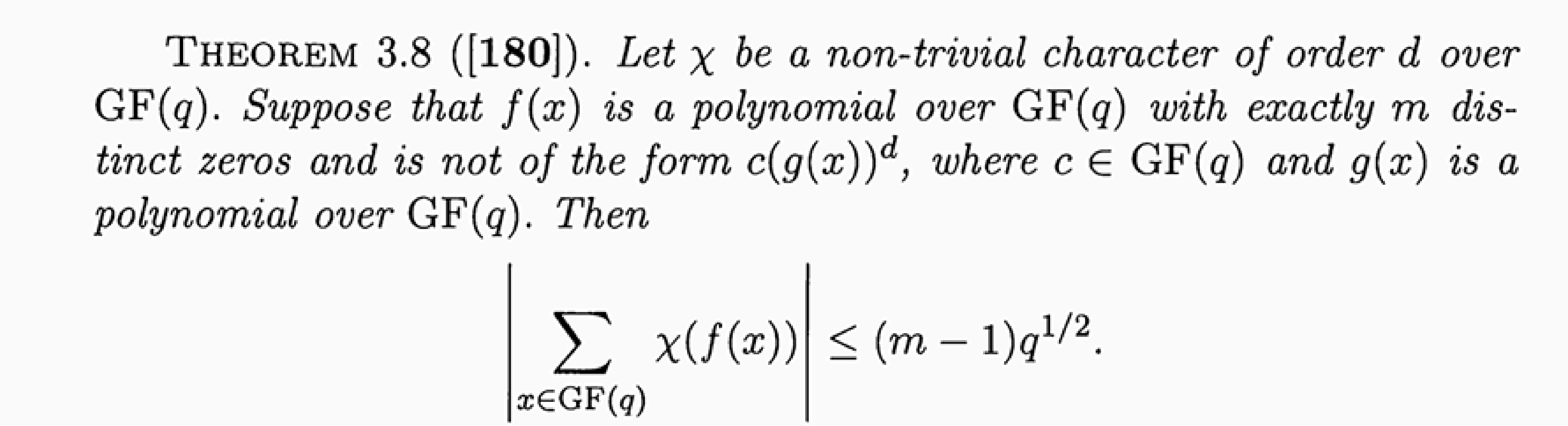


it follows that $s(u,v) = \frac{1}{2}(q-3)$. Hence,

$$\begin{aligned} \sum_{u,v\in V(G_n)} \left| s(u,v) - \frac{q}{2} \right| &= \sum_{u,v\in V(G_n)} \left| \frac{1}{2}(q-3) - \frac{q}{2} \right| \\ &= \binom{q}{2}\frac{3}{2} = \frac{3\left(q^2-q\right)}{4} \\ &= o(q^3). \quad \square \end{aligned}$$



THEOREM 3.7 ([**29, 38**]). *If $q > n^2 2^{2n-2}$, then P_q is n-e.c.*



THEOREM 3.8 ([**180**]). *Let χ be a non-trivial character of order d over* GF(q). *Suppose that $f(x)$ is a polynomial over* GF(q) *with exactly m distinct zeros and is not of the form $c(g(x))^d$, where $c \in$* GF(q) *and $g(x)$ is a polynomial over* GF(q). *Then*

$$\left| \sum_{x\in \mathrm{GF}(q)} \chi(f(x)) \right| \leq (m-1)q^{1/2}.$$

Another way to construct n.e.c graph

Affine Plane

n affine plane is a type of incidence geometry that abstracts the familiar properties of the Euclidean plane into a finite or infinite structure.

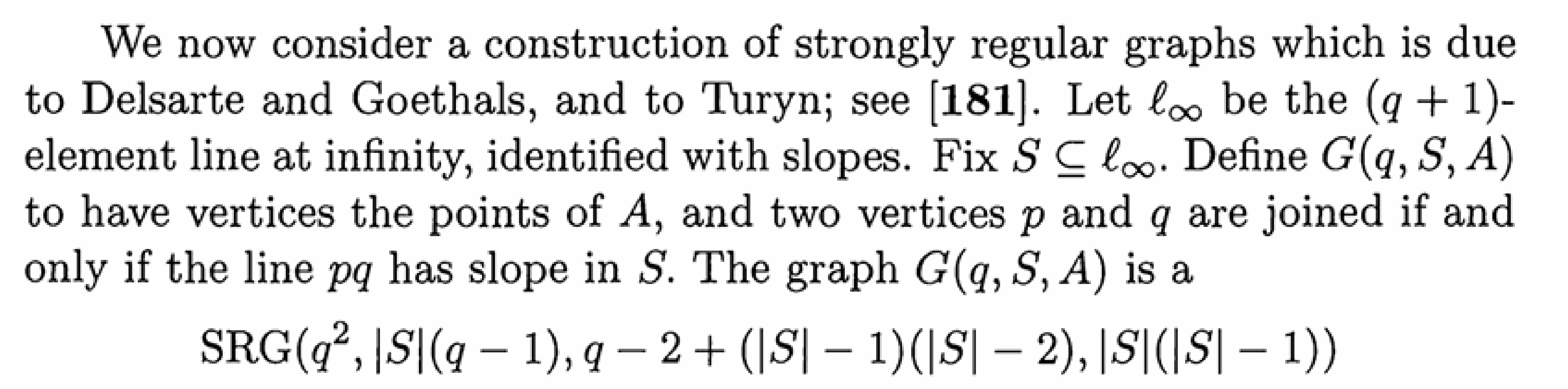
Formally, an affine plane is defined as a set of points and lines along with an incidence relation (which points lie on which lines) that satisfies the following axioms:

Unique Line Through Two Points:

Parallel Postulate (Existence and Uniqueness of Parallel Lines):

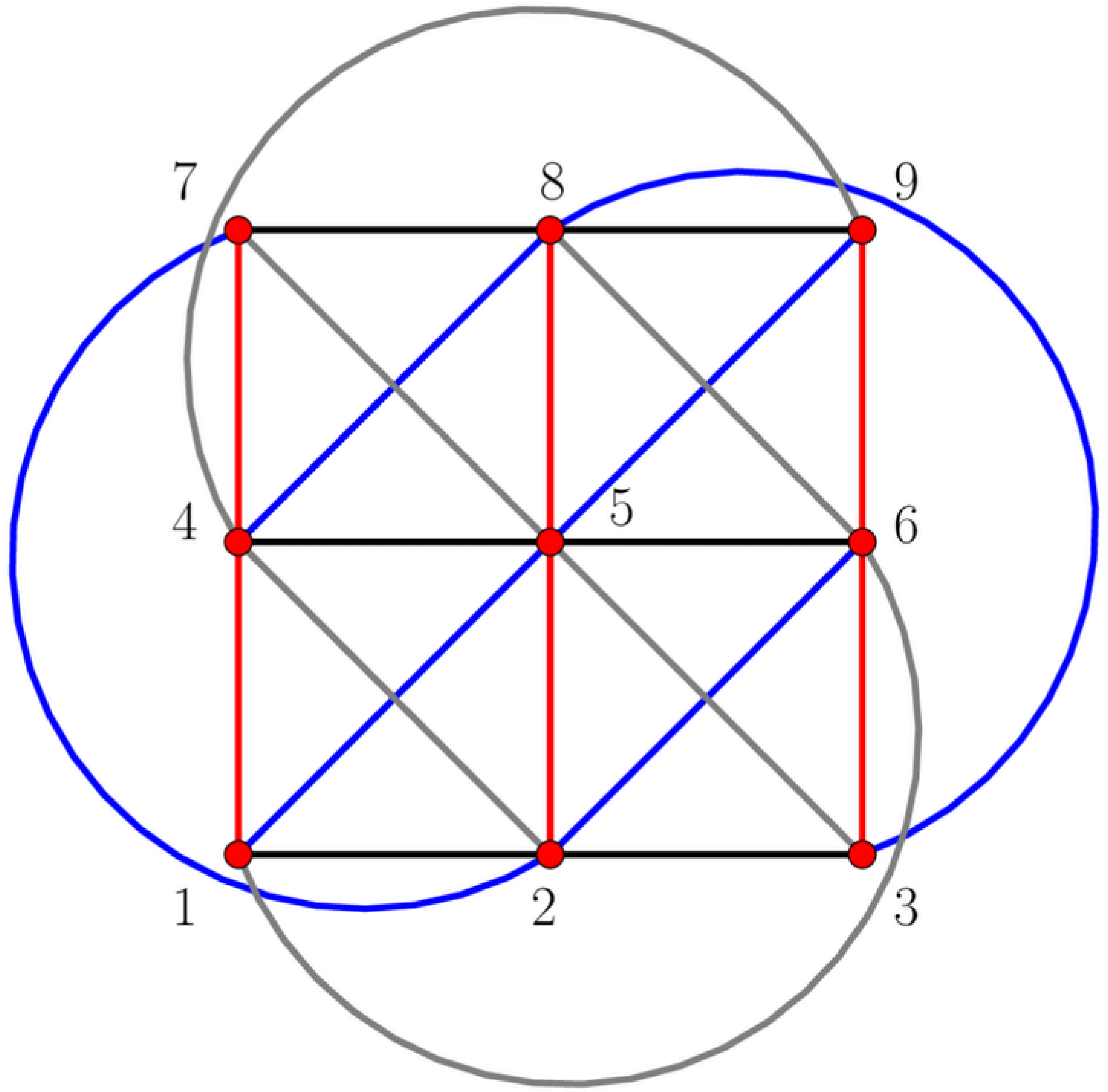
Existence of Non-Collinear Points (Non-Degeneracy):

Another way to construct n.e.c graph

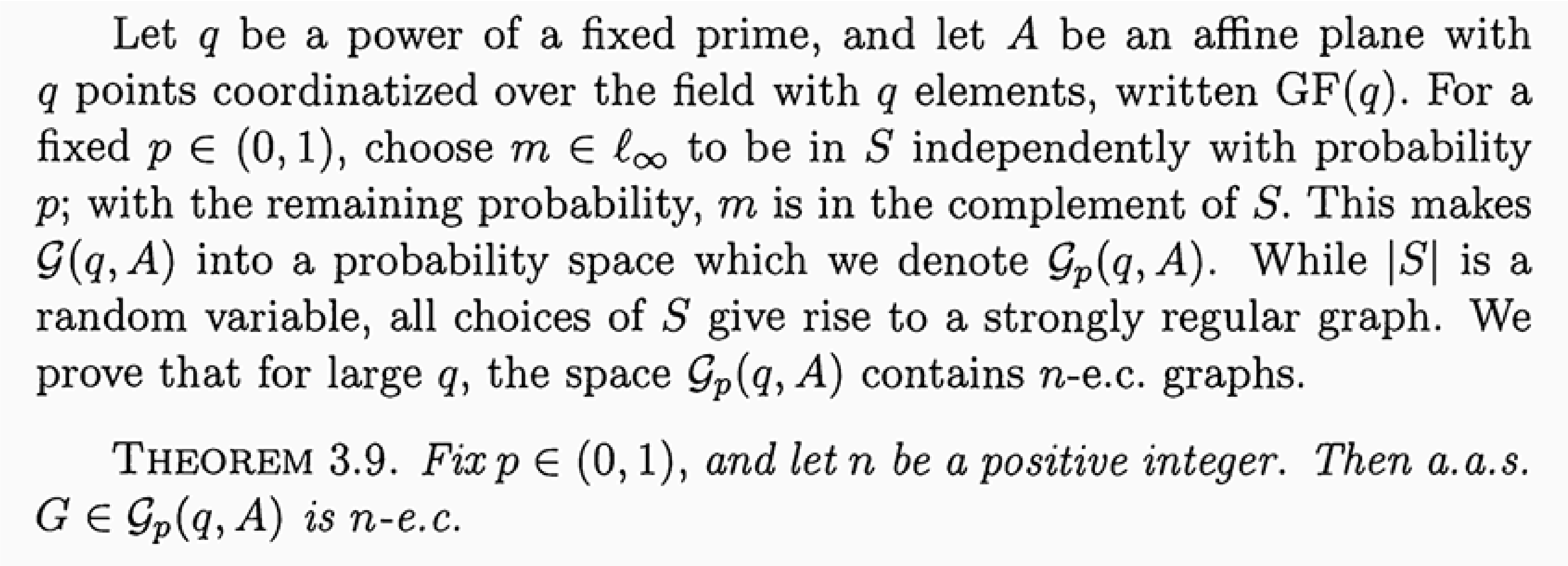


We now consider a construction of strongly regular graphs which is due to Delsarte and Goethals, and to Turyn; see [**181**]. Let ℓ_∞ be the $(q+1)$-element line at infinity, identified with slopes. Fix $S \subseteq \ell_\infty$. Define $G(q, S, A)$ to have vertices the points of A, and two vertices p and q are joined if and only if the line pq has slope in S. The graph $G(q, S, A)$ is a

$$\mathrm{SRG}(q^2, |S|(q-1), q-2+(|S|-1)(|S|-2), |S|(|S|-1))$$



Another way to construct n.e.c graph



Let q be a power of a fixed prime, and let A be an affine plane with q points coordinatized over the field with q elements, written $\mathrm{GF}(q)$. For a fixed $p \in (0,1)$, choose $m \in \ell_\infty$ to be in S independently with probability p; with the remaining probability, m is in the complement of S. This makes $\mathcal{G}(q,A)$ into a probability space which we denote $\mathcal{G}_p(q,A)$. While $|S|$ is a random variable, all choices of S give rise to a strongly regular graph. We prove that for large q, the space $\mathcal{G}_p(q,A)$ contains n-e.c. graphs.

THEOREM 3.9. *Fix $p \in (0,1)$, and let n be a positive integer. Then a.a.s. $G \in \mathcal{G}_p(q,A)$ is n-e.c.*

Citation Cartels

Citation cartels are groups of academic authors, journal editors, or publishers who collude to artificially inflate citation counts for their publications.
This practice is aimed at boosting individual or journal impact metrics, such as the Journal Impact Factor, which is often used to measure scientific influence and prestige.
In the current world

The ever-increasing competitiveness in the academic publishing market incentivizes journal editors to pursue higher impact factors

This fixation on higher impact factors leads some journals to artificially boost impact factors through the coordinated effort of a “citation cartel” of journals.

Citation Cartels

GLAD

GLAD (Global and Local Anomaly Detection) identifies citation cartels by combining global and local network anomaly scores, effectively spotting both dense citation loops and subtle collusive behaviors

CIDRE

CIDRE (Citation Detection and Reporting Engine) is a framework designed to detect abnormal citation patterns by analyzing citation graphs and spotting unusual clusters or bidirectional citation spikes.

ACTION

ACTION is a framework that models citation behaviors as probabilistic actions over time, aiming to detect coordinated citation boosts by examining temporal citation bursts and mutual citation sequences.

CIDRE

1. Graph Structure

It builds a directed weighted graph:

- Nodes: represent venues (journals/conferences), or sometimes individual authors or papers.
- Edges: a directed edge from node A to node B means A cites B.
- Weights: number of citations from A to B.

2. Mutual Citation Detection

Mutual Citation Ratio (MCR):

$MCR(A,B)=min(wAB,wBA)/max(wAB,wBA)$

- Where wAB is the number of times A cites B.
- High MCR close to 1 means symmetric mutual citations, which might be suspicious

3. Anomaly Detection

It uses a greedy clustering algorithm:

- Start with a suspicious edge (with high mutual citation).
- Expand the cluster by adding nodes that are highly connected with current cluster members.
- Measure cluster quality using metrics like Internal citation density, Contrast with citations to/from outside the cluster

THANK YOU