**T-79.7003: Graphs and Networks** **Fall 2013**

# Lecture 1: September 13

*Lecturer: Charalampos E. Tsourakakis* *Introduction*

## 1.1 Administrivia

**General Information:** Welcome to T-79.7003! Please call me Babis, which is a nickname for my first name Charalampos. If you have not registered but you are interested into attending the lectures, please contact me. This class will focus on the actively growing field of *network science*. One of the main goals of this field is to understand the graph structure and how it affects the diffusion of ideas, diseases and behaviors, as well as the performance of protocols and algorithms on networks. In general, complexity in social, biological and economical systems arises through pairwise interactions. Therefore, there exists a surging interest in understanding networks. This is not an easy task in general.

In this class we will focus on two aspects of network science. Specifically, the first part of the class will focus on *stochastic graph models*. The goal of this part is to go over standard probabilistic and discrete math techniques which are frequently used to analyze random graph models. We will study some basic properties of the Erdös-Rényi graph model. Then, we will study certain random graph models which reproduce some of the properties we observe in real-world networks. My plan is to go over preferential attachment, the Watts-Strogatz small-world model and Kleinberg's navigability result.

The second part of the class will focus on a major problem of algorithmic graph theory: *graph partitioning*. A major problem in network science is the detection of communities. Loosely speaking, a community is a set of vertices which is densely intra-connected but poorly connected to the rest of the graph. In this part of the class, we will go over Cheeger's inequality, spectral partitioning and one of the most significant conceptual advances of computer science: *every graph is essentially sparse*. It is worth outlining that this part of the class is important for a wide variety of network problems such as computing maximum flows and minimum cuts, solving linear systems of the form $Lx = b$ where $L$ is a Laplacian, but also for learning problems such as learning from labeled data on a graph or fitting a graph to cloud of points.

**Textbooks:** There is no required textbook. A suggested set of books available online and notes is available in the class Web page. I will provide you with lecture notes and slides. If you are interested in learning more, you are urged to read the references therein.

Here, is a list of books which are related to our material. I have annotated with magenta color the books/notes from which I plan to draw material.

- Random graphs, by *Béla Bollobás* [Bollobás, 2001]
- Complex graphs and networks, by *Fan Chung Graham and Linyuan Lu* [Chung and Lu, 2006]
- Random graphs, by *Svante Janson, Tomasz Luczak and Andrzej Rucinski* [Janson et al., 2000]
- Random Graph Dynamics, by *Rick Durrett* [Durrett, 2007]
- Random Graphs and Complex Networks, by *Remco Van Der Hofstad* available online at `http://www.win.tue.nl/~rhofstad/NotesRGCN2013.pdf`

- Networks, Crowds, and Markets: Reasoning About a Highly Connected World, by *David Easley and Jon Kleinberg* available online at `http://www.cs.cornell.edu/home/kleinber/networks-book/`
- Alan Frieze's notes, available online at `http://www.math.cmu.edu/~af1p/Teaching/RandomGraphs/RandomGraphs.html`

**Prerequisites:** I will assume familiarity with undergraduate material in algorithms, discrete mathematics, linear algebra and probability. Also, you should be familiar with a programming language such as C/C++, Matlab, Java, Python, Maple, Mathematica etc.

**Grading:** There will be three (3) homeworks. These homeworks are worth 15% of your final grade, 5% each. Most of the exercises will be theoretical, but there will also be programming exercises. You can use any programming language you want. In the class Web page, you will find a set of links to existing software that you may use for the homework purposes. The theoretical exercises will prepare you as well for the two exams that you will take. Each exam accounts of 30% of your final grade. Finally, you will work on a project that will account of 25% of your grade. There will be a list of suggested projects, but you are welcome to choose your own, as long as it is related to the class.

**Projects:** Projects will be consist of a project report and a presentation in class. Students who work on theoretical computer science/discrete mathematics are welcome to read a paper and present it. In case that after you choose your favorite paper you find it hard to understand, I will be happy to discuss it with you beforehand. Students who do research in data mining are welcome to conduct an experimental project. For the latter type of projects, collaboration in groups of two is welcomed. If you want to collaborate with more than one person, please contact me first.

**Homework policy:** The policy which I am going to adhere strictly to, is the following: you may discuss the problems with other students but you must write your solutions on your own and list your collaborators. You may not search the Web for solutions. You may consult outside materials, but you must cite your sources in your solutions. Please, do not bring me in an awkward position.

**Web site:** The class Web page can be found at:
`http://www.math.cmu.edu/~ctsourak/t797003-graphs-and-networks.html`.

**Resources:** In the class Web page you can find links to related classes, datasets, software and visualization software.

**Office hours:** Feel free to send me an email to arrange an appointment.

**Tentative plan:** For those of you who are currently in the dilemma of registering or not, here is roughly a sketch of the lectures so that you can make up your minds.

- Lecture 1 (13/9): Introduction, $G(n,p)$ and $G(n,m)$, asymptotic equivalence
- Lecture 2 (20/9): Subgraphs and Connectivity of $G(n,p)$

- Lecture 3 (27/9): Evolution of $G(n, p)$
- Lecture 4 (4/10): Maximum clique and chromatic number of $G(n, p)$
- Lecture 5 (11/10): Eigenvalues of G(n,p), detecting a hidden clique of size $O(\sqrt{n})$ in $G(n, 1/2)$
- Lecture 6 (18/10): Exam 1
- Lecture 7 (25/10): Preferential attachment (degree sequence), Watts-Strogatz model, Kleinbergs navigability result
- Lecture 8 (1/11): Conductance, Cheeger's inequality
- Lecture 9 (8/11): Every graph is essentially sparse I (preliminaries: graphs as electrical networks, random walks, effective resistance)
- Lecture 10 (15/11): Every graph is essentially sparse II (Spielman-Srivastava)
- Lecture 11 (22/11): Project presentations
- Lecture 12 (29/11): Project presentations
- Lecture 13 (6/12): Final exam

## 1.2 Real-world networks

The term *real-world networks* is used to refer to networks that appear in nature and society. Here are some examples of important real-world networks, see also [Albert and Barabási, 2002]:

- Human brain: our brain consists of neurons which form a network. The number of neurons in the region of human cortex is estimated to be $\sim 10^{10}$. Neurons are connected through synapses. The strength of synapses varies. But the average number of synapses of each neuron is in the range of 24 000-80 000 range for humans [Valiant, 2005].
- World Wide Web : Jim Gray in his 1998 Turing award address mentioned "The emergence of 'cyberspace' and the World Wide Web is like the discovery of a new continent". The vertices of the World Wide Web (WWW) are Web pages and the edges are hyperlinks (URLs) that point from one page to another. In 2011, Google reported that WWW has more than a trillion of edges.
- Internet: There exist two different types of Internet graphs, depending on the level we look at it. In the first type, vertices are routers and edges correspond to physical connections. The second level is the autonomous systems level, where a single vertex represents a domain, namely multiple routers and computers. An edge is drawn if there is at least one route that connects them.

  The first type of topology can be studied with the *traceroute* tool, whereas the second with *BGP tables*.

- Social and online social networks: Each vertex represents a human. Each edge corresponds to some sort of interaction, e.g., friendship, sexual interaction. Edges can be undirected or directed. For instance, an undirected edge if two Facebook accounts are connected or a directed edge if account $i$ follows account $j$ in Twitter. Nowadays, online social networks and social media are a part of our daily lives which can affect society immensely.

- Collaboration networks: There exist many types of collaboration networks. Vertices can represent for instance mathematicians or actors. An edge between two vertices is drawn when the mathematicians have co-authored a paper or when the actors played in the same movie respectively. Take a look at two famous examples of collaboration networks the Erdös collaboration network `http://www.oakland.edu/enp/` and the Kevin Bacon project `http://oracleofbacon.org/`.

- Food networks: Here, vertices are species and edges correspond predator-prey relations among them.

- Protein interaction networks: Vertices are proteins. Proteins $i$ and $j$ are connected if the bind together in order to carry out their biological function.

- Phone call networks: Here, vertices are phone numbers and each directed edge $(i, j)$ corresponds to a completed call, directed from the caller to the receiver.

- Wireless sensor networks: Vertices are autonomous sensors. These sensors monitor physical or environmental conditions, such as temperature, brightness, humidity etc. A directed edge $(i, j)$ suggests that information can be passed from sensor $i$ to sensor $j$.

- Power grid: Here, vertices are generators, transformers and substations. Edges correspond to high-voltage transmission lines.

- Financial networks: Here, vertices are financial institutions. Edges can correspond to different types of interactions, for instance borrower-lender relations.

- Blog networks: Vertices represent blogs. An edge $(i, j)$ exists if the $i$-th blog has blog $j$ in its blogroll or a URL pointing to $j$.

There exists another important conceptual category of networks. These are networks that are created after processing other types of datasets, e.g., a cloud of points in $\mathbb{R}^d$. For instance, gene co-expression networks are created from microarray datasets [Zhang and Horvath, 2005].

## 1.3 Empirical properties of real-world networks

We saw a diverse list of real-world networks. Do they share any common characteristics? The answer is yes. This fact is very surpising if you think about it for a while. Why should Of course, we have to outline that the fact that they share some common characteristic does not imply that they have the same graph structure. We discussed in class the following:

- Static networks
    1. Heavy tails
    2. Clustering coefficients

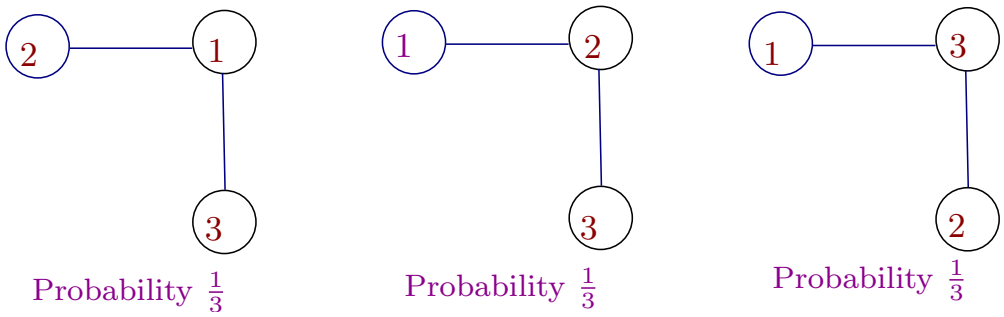


Figure 1.1: A random graph on $\{1, 2, 3\}$ with 2 edges with the uniform distribution

3. Communities
4. Small diameters
5. Eigenvalues

- Time-evolving networks
  1. Densification
  2. Shrinking diameters
- Web graph
  1. Bow-tie structure
  2. Bipartite cliques

## 1.4 Random Graphs

### 1.4.1 What is a random graph?

Formally, when we are given a graph $G$ and we say this is a random graph, we are wrong. A given graph is fixed, there is nothing random to it. What we mean though through this term abuse is that this graph was sampled out of a set of graphs according to a probability distribution. For instance, Figure 1.1 shows the three possible graphs on vertex set $[3] = \{1, 2, 3\}$ with 2 edges. The probability distribution is the uniform, namely, each graph has the same probability $\frac{1}{3}$ to be sampled.

### 1.4.2 $G(n, p), G(n, m)$

- *Random binomial graphs,* $G(n, p)$: This model has two parameters, the number of vertices $n$ and a probability parameter $0 \leq p \leq 1$. Let $\mathcal{G}$ be the family of all possible labelled graphs on the vertex set $[n]$. Notice $|\mathcal{G}| = 2^{\binom{n}{2}}$. The $G(n, p)$ model assigns to a graph $G \in \mathcal{G}$ the following probability

$$\mathbf{Pr}[G] = p^{|E(G)|}(1-p)^{\binom{n}{2}-|E(G)|}.$$

- *Uniform random graph,* $G(n, m)$: This model has two parameters, the number of vertices $n$ and the number of edges $m$, where $0 \leq m \leq \binom{n}{2}$. This model assigns to all labelled graphs on the vertex set $[n]$ with exactly $m$ edges equal probability. In other words,

$$\mathbf{Pr}[G] = \begin{cases} \frac{1}{\binom{\binom{n}{2}}{m}} & \text{if } |E(G)| = m \\ 0 & \text{if } |E(G)| \neq m \end{cases}$$

Notice that in the $G(n,p)$ model we toss a coin independently for each edge, and with probability $p$ we add it to the graph. In expectation there will be $p\binom{n}{2}$ edges. When $p = \frac{m}{\binom{n}{2}}$, then a random binomial graph in expectation has $m$ edges and intuitively the two models should behave similarly. For this $p$ the two models behave similarly in a quantifiable sense. We start with the following simple observation.

**Fact 1.1** *A random graph $G(n,p)$ with $m$ edges is equally likely to be any of the $\binom{\binom{n}{2}}{m}$ graphs with $m$ edges.*

**Proof:**

Consider any graph with $m$ edges, call it $G$.

$$\begin{aligned} \mathbf{Pr}[G(n,p) = G || E(G(n,p))| = m] &= \frac{\mathbf{Pr}[G(n,p) = G]}{\mathbf{Pr}[|E(G(n,p))| = m]} \\ &= \frac{p^m(1-p)^{\binom{n}{2}-m}}{\binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m}} \\ &= \frac{1}{\binom{\binom{n}{2}}{m}} \end{aligned}$$

■

**Definition 1.2** *Define a graph property $\mathcal{P}$ as a subset of all possible labelled graphs. Namely $\mathcal{P} \subseteq 2^{\binom{[n]}{2}}$.*

For instance $\mathcal{P}$ can be the set of planar graphs or the set of graphs that contain a Hamiltonian cycle. We will call a property $\mathcal{P}$ as monotone increasing if $G \in \mathcal{P}$ implies $G + e \in \mathcal{P}$. For instance the Hamiltonian property is monotone increasing. Similarly, we will call a property $\mathcal{P}$ as monotone decreasing if $G \in \mathcal{P}$ implies $G - e \in \mathcal{P}$. For instance the planarity property is monotone decreasing.

**Exercise:** Think of other monotone increasing and decreasing properties.

Consider any monotone increasing property $\mathcal{P}$. Intuitively, the more edges the graph has, the more likely a random graph has property $\mathcal{P}$[1]. Indeed,

**Lemma 1.3** *Suppose $\mathcal{P}$ is a monotone increasing property and $0 \leq p_1 < p_2 \leq 1$. Let $G_i \sim G(n,p_i), i = 1,2$. Then,*

$$\mathbf{Pr}[G_1 \in \mathcal{P}] \leq \mathbf{Pr}[G_2 \in \mathcal{P}].$$

**Proof:** We will generate $G_2 \sim G(n,p_2)$ from a graph $G_1 \sim G(n,p_1)$. The idea is called *coupling*. After generating $G_1$ we will generate a graph $G \sim G(n,p)$ and we will output the union of $G_1 \cup G$ as our $G_2$. We

[1] I will use interchangeably the terms a graph *has* property $\mathcal{P}$ and a graph *belongs* in $\mathcal{P}$.

need to choose $p$ in such way that we respect the probability distributions. To see how to choose $p$ observe the following: an edge in $G_2$ does not exist with probability $(1-p_2)$. In $G_1 \cup G$ this happens with probability $(1-p)(1-p_1)$. By setting

$$(1-p_2) = (1-p)(1-p_1)$$

and solving for $p$ we have achieved our goal. Given that the property is monotone increasing, we obtain the result. ■

**Exercise:** Prove an analog lemma for the $G(n,m)$ model.

Now we prove two facts before we give a general statement for the asymptotic equivalence of the two models.

**Fact 1.4** *Let $\mathcal{P}$ be any graph property, $p = \frac{m}{\binom{n}{2}}$, where $m = m(n)$, $\binom{n}{2} - m \to +\infty$. Then, asymptotically*

$$\mathbf{Pr}\left[G(n,m) \in \mathcal{P}\right] \leq \sqrt{2\pi m}\mathbf{Pr}\left[G(n,p) \in \mathcal{P}\right].$$

**Proof:**

The probability that we obtain a given graph $G$ depends only on the number of its edges. Also notice that there exist $\binom{\binom{n}{2}}{k}$ graphs with $k$ distinct edges, for any $0 \leq k \leq \binom{n}{2}$. Therefore, from the law of total probability we obtain the following expression:

$$\begin{aligned}
\mathbf{Pr}\left[G(n,p) \in \mathcal{P}\right] &= \sum_{m'=0}^{\binom{n}{2}} \mathbf{Pr}\left[|E(n,p)| = m'\right] \times \mathbf{Pr}\left[G(n,p) \in \mathcal{P} || E(n,p)| = m'\right] \\
&\geq \mathbf{Pr}\left[|E(n,p)| = m\right] \times \mathbf{Pr}\left[G(n,p) \in \mathcal{P} || E(n,p)| = m\right] \\
&= \mathbf{Pr}\left[|E(n,p)| = m\right] \times \mathbf{Pr}\left[G(n,m) \in \mathcal{P}\right].
\end{aligned}$$

It suffices to prove that

$$\mathbf{Pr}\left[|E(n,p)| = m\right] \geq \frac{1}{\sqrt{2\pi m}}.$$

For this purpose we will use Stirling's formula[2]

$$n! = (1+o(1))\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}.$$

Also, we observe that the random variable $|E(n,p)|$ is a binomial variable, i.e., $|E(n,p)| \sim Bin(\binom{n}{2}, p)$. Therefore,

$$\begin{aligned}
\mathbf{Pr}\left[|E(n,p)| = m\right] = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m} &\approx \left(\frac{\binom{n}{2}}{2\pi m\left(\binom{n}{2} - m\right)}\right)^{1/2} \\
&\geq \frac{1}{\sqrt{2\pi m}}.
\end{aligned}$$

[2]For those of you that think this formula is magic or have forgotten how it is proved, check out this post `http://gowers.wordpress.com/2008/02/01/removing-the-magic-from-stirlings-formula/` by Timothy Gowers.

■

**Exercise:** The following fact is left as an exercise. You can solve it either by using the central limit theorem or by more tedious computations using appropriate asymptotic approximations.

**Fact 1.5** *Let $\mathcal{P}$ be a monotonically increasing (decreasing) graph property, $p = \frac{m}{\binom{n}{2}}$. Then, asymptotically*

$$\mathbf{Pr}\left[G(n,m) \in \mathcal{P}\right] \leq 3\mathbf{Pr}\left[G(n,p) \in \mathcal{P}\right].$$

The following theorem gives precise conditions for the asymptotic equivalence of $G(n,p), G(n,m)$ [Frieze and Karoński, ].

**Theorem 1.6** *Let $0 \leq p_0 \leq 1, s(n) = n\sqrt{p(1-p)} \to +\infty$, and $\omega(n) \to +\infty$ as $n \to +\infty$. Then,*

*(a) if $\mathcal{P}$ is any graph property and for all $m \in \mathbb{N}$ such that $|m - \binom{n}{2}p| < \omega(n)s(n)$, the probability $\mathbf{Pr}\left[G(n,m) \in \mathcal{P}\right] \to p_0$, then $\mathbf{Pr}\left[G(n,p) \in \mathcal{P}\right] \to p_0$ as $n \to +\infty$.*

*(b) if $\mathcal{P}$ is a monotone graph property and $p_- = p_0 - \frac{\omega n s(n)}{n^3}$, $p_+ = p_0 + \frac{\omega n s(n)}{n^3}$ then from the facts that $\mathbf{Pr}\left[G(n,p_-) \in \mathcal{P}\right] \to p_0, \mathbf{Pr}\left[G(n,p_+) \in \mathcal{P}\right] \to p_0$, it follows that $\mathbf{Pr}\left[G(n,p\binom{n}{2}) \in \mathcal{P}\right] \to p_0$ as $n \to +\infty$.*

### 1.4.3 History

The theory of random graphs was founded by Paul Erdös and Alfred Rényi in a series of seminal papers. Erdös and Rényi studied originally the $G(n,m)$ model. Gilbert proposed the $G(n,p)$ model. Some people refer to random binomial graphs as Erdös-Rényi or Erdös-Rényi-Gilbert. Nonetheless, it was Erdös and Rényi who set the foundations of modern random graph theory.

Before the series of Erdös-Rényi papers, Erdös had discovered that the probabilistic method could be used to tackle problems whose statements were purely deterministic. For instance, one of the early uses of random graphs was in Ramsey theory. We define the Ramsey number

$$R(k,l) = \min\{n : \forall c : E(K_n) \to \{\text{red,blue}\} \exists \text{ red } K_k \text{ or blue } K_l\}.$$

**Example:** Prove $R(3,3) = 6$. The next challenge is to show $R(4,4) = 18$.

In one of the next lectures we will study the maximum clique in $G(n,p)$. Specifically, by studying the maximum clique size in $G(n,1/2)$, we will see why $R(k,k) \geq 2^{k/2}$. Now, let's see a proof based on the union bound.

**Theorem 1.7 (Erdös, 1947)**

$$R(k,k) \geq 2^{k/2}.$$

**Proof:** Color each edge of the complete graph $K_n$ with red or blue by tossing a fair coin, independently from the other edges. For a fixed subset $S \subseteq [n]$, $|S| = k$ let $A_S$ be the event that $S$ is monochromatic, i.e., all the $\binom{k}{2}$ edges get the same color. Clearly, $\mathbf{Pr}\left[A_S\right] = 2^{1-\binom{k}{2}}$. Notice that if $\mathbf{Pr}\left[\cup_{S\subseteq V, |S|=k} A_S\right] < 1$ then the probability that none of the $k$-sets is monochromatic is $> 0$ which means that there exists a 2-coloring which violates the Ramsey property. Hence this would suggest that $R(k,k) > n$.

Based on the union bound

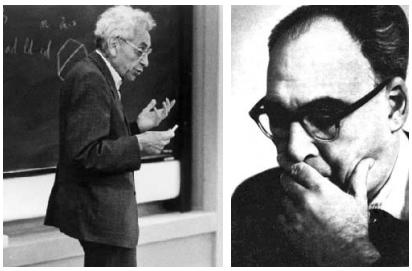

Figure 1.2: Erdös & Rényi, founders of random graph theory

$$\cup_{S\subseteq V,|S|=k}\mathbf{Pr}\left[A_S\right] \leq \binom{n}{k}2^{1-\binom{k}{2}}$$

we can deduce that $R(k,k) > n$ if $\binom{n}{k}2^{1-\binom{k}{2}} < 1$. When $n = \lfloor 2^{k/2} \rfloor$ then this condition holds. Let's check it.

$$\binom{n}{k}2^{1-\binom{k}{2}} < \frac{n^k}{k!}2^{1-\binom{k}{2}} < 1.$$

■

## 1.5 First and Second Moment Method

We will end our first lecture with two elementary probabilistic tools, which are very powerful. Just with these tools, many non-trivial results can be proved.

**Theorem 1.8 (First Moment Method)** *Let $X$ a be non-negative integer valued random variable. Then,*

$$\mathbf{Pr}\left[X > 0\right] \leq \mathbb{E}\left[X\right].$$

Sometimes, it is much easier to compute $\mathbb{E}\left[X\right]$ than $\mathbf{Pr}\left[X > 0\right]$. Therefore, if we want to prove that $\mathbf{Pr}\left[X > 0\right]$ tends to 0 as $n \to +\infty$ it suffices to show that $\mathbb{E}\left[X\right] = o(1)$.

However, if we want to show that $\mathbf{Pr}\left[X > 0\right]$ we cannot use the First Moment Method. The use of the following inequality is also known as Chebyshev's inequality.

**Theorem 1.9 (Second Moment Method)** *If $X$ is a nonnegative integer valued random variable then*

$$\mathbf{Pr}\left[X > 0\right] \geq 1 - \frac{\mathbb{V}ar\left[X\right]}{(\mathbb{E}\left[X\right])^2}.$$

Typically, Theorem 1.9 will do the work. But sometimes a stronger form of the second moment method is needed. It is stated as the following theorem.

**Theorem 1.10 (Strong Second Moment Method)** *If $X$ is a nonnegative integer valued random variable then*

$$\mathbf{Pr}\left[X > 0\right] \geq 1 - \frac{\mathbb{V}ar\left[X\right]}{\mathbb{E}\left[X^2\right]}.$$

**Exercise:** Prove Theorems 1.8, 1.9 and 1.10.

# References

[Albert and Barabási, 2002] Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

[Bollobás, 2001] Bollobás, B. (2001). *Random graphs*, volume 73. Cambridge university press.

[Chung and Lu, 2006] Chung, F. R. K. and Lu, L. (2006). *Complex graphs and networks*. Number 107. AMS Bookstore.

[Durrett, 2007] Durrett, R. (2007). *Random graph dynamics*, volume 20. Cambridge university press.

[Frieze and Karoński, ] Frieze, A. and Karoński, M. Introduction to random graphs.

[Janson et al., 2000] Janson, S., Luczak, T., and Rucinski, A. (2000). *Random graphs*. Cambridge Univ Press.

[Valiant, 2005] Valiant, L. G. (2005). Memorization and association on a realistic neural model. *Neural computation*, 17(3):527–555.

[Zhang and Horvath, 2005] Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128.