Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Εκτίμηση Παραμέτρων Γραφημάτων από Θορυβώδη Δείγματα και Ερωτήματα

Διπλωματικη Εργασια

Αθανάσιος Πήττας

Επιβλέπων: Δημήτρης Φωτάκης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εργαστηριο Λογικης και Επιστημης Υπολογισμων
Αθήνα, Οκτώβριος 2019

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Εκτίμηση Παραμέτρων Γραφημάτων από Θορυβώδη Δείγματα και Ερωτήματα

Διπλωματικη Εργασια

Αθανάσιος Πήττας

Επιβλέπων: Δημήτρης Φωτάκης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Οκτωβρίου 2019.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτρης Φωτάκης	Άρης Παγουρτζής	Νικόλαος Παπασπύρου
Αν. Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Εργαστήριο Λογικής και Επιστήμης Υπολογισμών
Αθήνα, Οκτώβριος 2019

(Υπογραφή)

.....
Αθανάσιος Πήττας
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η μέτρηση του πλήθους εμφανίσεων μικρών υπογραφημάτων σε μεγαλύτερα γραφήματα είναι μια κρίσιμη εργασία για πολλές εφαρμογές. Πιο συγκεκριμένα, τα τρίγωνα εμπλέκονται στον ορισμό μετρικών όπως ο συντελεστής ομαδοποίησης (clustering coefficient) και ο λόγος μεταβατικότητας (transitivity ratio), που είναι θεμελιώδεις στη μελέτη σύνθετων δικτύων. Μια δυσκολία στον υπολογισμό αυτών των παραμέτρων είναι ότι συχνά τα γραφήματα προς μελέτη δεν είναι πλήρως διαθέσιμα για διάφορους λόγους.

Σε αυτή την διπλωματική εργασία, ορίζουμε ένα μοντέλο θορύβου παρακινούμενοι από την ιδιωτικότητα στα κοινωνικά δίκτυα, δηλαδή το γεγονός ότι οι χρήστες μπορούν να θέτουν τους φίλους τους ως κρυφούς. Ειδικότερα, έχουμε ένα αυθεντικό γράφημα όπου κάθε κορυφή εκτελεί ένα πείραμα Bernoulli με κάποια γνωστή πιθανότητα επιτυχίας και το θορυβώδες δείγμα ορίζεται αφαιρώντας τις ακμές για τις οποίες και τα δύο άκρα σημειώνουν επιτυχία στα πειράματα. Ο στόχος είναι η εκτίμηση του πλήθους ακμών και τριγώνων του αυθεντικού γραφήματος εντός μικρού πολλαπλασιαστικού σφάλματος, με μεγάλη πιθανότητα. Αρχικά, βρίσκουμε τέτοιες εκτιμήτριες και αποδεικνύουμε αντίστοιχα κάτω φράγματα για το απαιτούμενο πλήθος δειγμάτων, βασισμένα στην θεωρία πληροφορίας. Έπειτα, επιτρέπουμε στους αλγορίθμους μας να έχουν επιπλέον πρόσβαση στο αυθεντικό γράφημα μέσω μαντείου που αποκαλύπτει την πραγματική γειτονιά κορυφών και επεκτείνουμε τις προηγούμενες εκτιμήτριες για αυτή την περίπτωση. Ο αριθμός των δειγμάτων που χρειάζονται γίνεται σταθερός και το πλήθος των ερωτημάτων προς το μαντείο εξαρτάται μόνο από τις παραμέτρους ακρίβειας και όχι από το μέγεθος του γραφήματος.

Λέξεις κλειδιά: Εκτιμητική, Τρίγωνα, Μάθηση, Θεωρία Πληροφορίας

Abstract

Counting the number of occurrences of fixed subgraphs in larger graphs is a crucial task for many applications. More specifically, triangles are involved in the definition of metrics such as the clustering coefficient and the transitivity ratio, which are fundamental in complex network analysis. A difficulty for the calculation of these parameters is that input graphs are often incomplete for various reasons.

In this thesis we introduce a noise model motivated by the privacy constraints in social networks, that is, the fact that users can mark their friends as hidden from the public. More precisely, we have an underlying graph where each vertex performs a Bernoulli trial with some known probability of success and the noisy sample graph is defined by removing the edges for which both endpoints succeed in the trials. The goal is to estimate the number of edges and triangles of the underlying graph within a small relative error, with high probability. First, we derive such estimators and prove matching information theoretic lower bounds for the sample complexity of these tasks. Then, we allow our algorithms to have additional query access to the underlying graph by asking vertices to reveal their true neighborhood and we extend the previous estimators in this setting. The number of samples required becomes constant and the number of queries depends only on the accuracy parameters and not the size of the graph.

Keywords: Estimation, Triangles, Learning, Information Theory

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τους καθηγητές του Εργαστηρίου Λογικής και Επιστήμης Υπολογισμών, κ. Ζάχο, κ. Παγουρτζή και κ. Φωτάκη. Είναι οι άνθρωποι που με δίδαξαν θεωρητική πληροφορική, ο καθένας με τον δικό του τρόπο, και με υποστήριξαν στις αποφάσεις μου τα τελευταία χρόνια των προπτυχιακών μου σπουδών. Ιδιαίτερα, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα κ. Φωτάκη για την καθοδήγηση που μου προσέφερε κατά την διεκπεραίωση αυτής της διπλωματικής και την προσπάθεια του να με εισάγει στην ερευνητική διαδικασία. Ο ενθουσιασμός που μεταδίδει για την θεωρητική πληροφορική ήταν από τα καλύτερα κίνητρα για να ασχοληθώ και εγώ με αυτά τα θέματα. Ευχαριστώ ακόμα τα υπόλοιπα μέλη του εργαστηρίου Στρατή, Αγγέλα, Παναγιώτη, Τζέλλα για το ευχάριστο κλίμα, την παρέα και τις επιστημονικές συζητήσεις. Τέλος, δε γίνεται να παραλείψω τους φίλους μου Κωνσταντίνα, Άγγελο, Γιάννη, Γιώργο, Αλεξία, Χάρις, Κωνσταντίνο, Θεώνη που ήταν δίπλα μου όλο αυτό τον καιρό.

Θανάσης

Στον παππού μου

Εκτεταμένη Ελληνική Περίληψη

Σε αυτό το κεφάλαιο παρουσιάζουμε συνοπτικά το περιεχόμενο και τα αποτελέσματα αυτής της διπλωματικής εργασίας. Αρχικά εισάγουμε το πρόβλημα της παρούσας εργασίας εξηγώντας το κίνητρο για τον ορισμό του. Έπειτα παρουσιάζουμε την συνεισφορά μας, όπου παραθέτουμε τα βασικά συμπεράσματα χωρίς τις αποδείξεις και τις τεχνικές λεπτομέρειες, οι οποίες υπάρχουν στο αγγλικό κείμενο.

Εισαγωγή

Ένα πολύ σημαντικό πρόβλημα στην στατιστική είναι η εκτίμηση μιας άγνωστης παραμέτρου κάποιας κατανομής. Για παράδειγμα, κάτι που μαθαίνουμε σε ένα εισαγωγικό μάθημα πιθανοτήτων είναι η εκτίμηση της μέσης τιμής μιας κανονικής κατανομής από δείγματά της. Το θέμα αυτής της διπλωματικής είναι η εκτίμηση παραμέτρων γραφημάτων από θορυβώδη δείγματά τους, κάτι που συνιστά σημαντικό θεωρητικό και πρακτικό πρόβλημα για τους λόγους που εξηγούνται παρακάτω.

Τα γραφήματα είναι δομές που συναντιούνται παντού για την αναπαράσταση δεδομένων, από την βιολογία μέχρι τα κοινωνικά δίκτυα. Σε αυτή την εργασία εστιάζουμε στα κοινωνικά δίκτυα [B+16] τα οποία είναι μη κατευθυνόμενα γραφήματα με μια κορυφή για κάθε χρήστη και μια ακμή για κάθε ζεύγος χρηστών που είναι φίλοι. Αυτά τα γραφήματα έχουν πολλές ενδιαφέρουσες ιδιότητες και έχουν μελετηθεί πολύ στη βιβλιογραφία [New03, AB02, DM13]. Επίσης το μέγεθος τους στη πράξη είναι τεράστιο, κάτι που γεννά την ανάγκη για αποδοτικούς αλγορίθμους.

Όμως τι θέλουμε να υπολογίζουμε με αλγορίθμους στα κοινωνικά δίκτυα; Τα τρίγωνα είναι σημαντικές δομές στα γραφήματα αυτά διότι γεννιούνται από φυσικούς μηχανισμούς. Πιο συγκεκριμένα, δύο άνθρωποι που έχουν πολλούς κοινούς φίλους, τείνουν να δημιουργούν δεσμό φιλίας και μεταξύ τους, κάτι που αναφέρεται ως *μεταβατικότητα*, και επιπλέον, άνθρωποι που είναι φίλοι τείνουν να έχουν πολλούς κοινούς φίλους γιατί πιθανότατα έχουν παρόμοια ενδιαφέροντα, κάτι που αναφέρεται ως *ομοφιλία* [WF+94]. Για αυτούς τους λόγους έχουν οριστεί μετρικές για να ποσοτικοποιήσουν την μεταβατικότητα και την ομοφιλία, οι οποίες είναι ο *λόγος μεταβατικότητας* [NWS02] και ο *συντελεστής ομαδοποίησης* [WS98]. Οι ορισμοί και των δύο βασίζονται στη μέτρηση τριγώνων.

Κίνητρο και Ορισμός του Μοντέλου

Αν και ένα μεγάλο μέρος της βιβλιογραφίας εστιάζει στο σχεδιασμό *γρήγορων* αλγορίθμων, ώστε να μπορούν να εκτελεστούν με είσοδο τεράστια δίκτυα, το σημείο αφετηρίας της παρούσας εργασίας είναι η *ιδιωτικότητα* στα κοινωνικά δίκτυα. Πιο συγκεκριμένα, κάθε χρήστης έχει την δυνατότητα να ορίσει τους φίλους του ως *κρυφούς*, καθιστώντας τα κοινωνικά δίκτυα ένα τέλειο παράδειγμα περιβάλλοντος με περιορισμένη πληροφορία. Ένα μοντέλο το οποίο περιλαμβάνει το χαρακτηριστικό που μόλις περιγράφηκε, δηλαδή τους κρυφούς φίλους, προτάθηκε στο [CEK+15]. Ο σκοπός σε εκείνη την εργασία ήταν να βρεθούν αποδοτικοί αλγόριθμοι για τον υπολογισμό ιδιοτήτων του δικτύου από τη σκοπιά κάθε χρήστη ξεχωριστά, με σεβασμό δηλαδή στην ιδιωτικότητα των άλλων χρηστών. Ορμώμενοι από αυτές τις ιδέες, πάμε ένα βήμα παραπέρα και ρωτάμε πως γίνεται βλέποντας το δημόσιο δίκτυο (δηλαδή αυτό που οι κρυφές σχέσεις μεταξύ φίλων είναι αόρατες) να εξάγουμε συμπεράσματα για το πραγματικό δίκτυο (που περιέχει κρυφές και μη κρυφές σχέσεις). Έτσι οδηγούμαστε στο εξής μοντέλο.

Στο μοντέλο μας υπάρχει το *αυθεντικό* γράφημα $G = (V, E)$ όπου V είναι το σύνολο των κορυφών και E το σύνολο των ακμών. Ένα *δείγμα* G_s του γραφήματος αυτού παράγεται από την εξής τυχαία διαδικασία. Κάθε κορυφή εκτελεί μια δοκιμή Bernoulli με πιθανότητα επιτυχίας p. Έστω $\{X_v\}_{v \in V}$ οι τυχαίες μεταβλητές για όλες τις κορυφές. Το G_s ορίζεται να έχει τις ίδιες κορυφές με το G και οι ακμές του ορίζονται

$$E_s = \{(u, v) \in E(G) \mid X_u = 0 \vee X_v = 0\}$$

Δηλαδή αυτό που περιγράφεται παραπάνω είναι ότι κάθε χρήστης με πιθανότητα p αποφασίζει αν θέλει να κρύψει τους γείτονές του και αν δύο φίλοι πάρουν αυτή την απόφαση η σχέση μεταξύ τους γίνεται αόρατη στο δημόσιο δείγμα. Μπορούμε να βλέπουμε το παραπάνω σαν τον ορισμό μιας κατανομής πιθανότητας πάνω στα γραφήματα n κορυφών, την οποία θα συμβολίζουμε με $\mathcal{P}_G$.

Θα επιτρέπουμε στους αλγορίθμους μας να έχουν επιπλέον πρόσβαση στο αυθεντικό γράφημα μέσω *μαντείου* το οποίο με είσοδο μια κορυφή v θα επιστρέφει την γειτονιά αυτής της κορυφής $\Gamma(v)$ στο αυθεντικό γράφημα.

Ο λόγος για το μαντείο είναι ότι κάθε δείγμα αποκαλύπτει πολλή πληροφορία, η οποία όμως μπορεί να είναι περιττή από ένα σημείο και μετά. Επίσης ένα ολόκληρο δείγμα μπορεί να είναι πολύ ακριβό σε σχέση με μια κλίση του μαντείου. Έτσι θέλουμε με λίγα στοχευμένα ερωτήματα στο μαντείο να αποσπάσουμε ισοδύναμη πληροφορία που θα προσφερόταν από ολόκληρα δείγματα.

Στόχος

Τα προβλήματα που εξετάζουμε σε αυτή την εργασία είναι αρχικά η εκτίμηση του πίνακα γειτνίασης του αυθεντικού γραφήματος και έπειτα η εκτίμηση του πλήθους ακμών και τριγώνων στο αυθεντικό γράφημα. Πιο συγκεκριμένα, έστω $G_1, \ldots, G_N$ τα δείγματα. Θέλουμε εκτιμήτριες που να πληρούν τα εξής.

1. **Εκτίμηση του πίνακα γειτνίασης**. Δεδομένου μιας παραμέτρου $\delta \in (0, 1]$ ο στόχος είναι να βρεθεί γράφημα $\hat{G}$ τέτοιο ώστε

$$\mathop{\mathbb{P}}_{G_1,\ldots,G_N \sim \mathcal{P}_G}(\hat{G} = G) \geq 1 - \delta \tag{1}$$

2. **Προσεγγιστική εκτίμηση ακμών**. Για κάθε $\varepsilon, \delta \in (0,1]$ ψάχνουμε μια εκτιμήτρια $\hat{m}$ για τις ακμές του αυθεντικού γραφήματος $m = |E(G)|$ έτσι ώστε

$$\mathop{\mathbb{P}}_{G_1,\ldots,G_N \sim \mathcal{P}_G}(|\hat{m} - m| \leq \varepsilon m) \geq 1 - \delta$$

3. **Προσεγγιστική εκτίμηση τριγώνων**. Το ίδιο για εκτίμηση τριγώνων.

Εκτίμηση από Δείγματα

Στη συνέχεια παρουσιάζουμε τα αποτελέσματά μας για τα παραπάνω προβλήματα. Αρχικά εξετάζουμε αλγορίθμους που έχουν πρόσβαση μόνο σε δείγματα του αυθεντικού γραφήματος και όχι πρόσβαση στο μαντείο. Αυτό θα δώσει ιδέες για την βελτίωσή τους στην επόμενη ενότητα όταν επιτρέψουμε χρήση του μαντείου.

Εκτίμηση του Αυθεντικού Γραφήματος

Αρχικά εξετάζουμε μια απλή εκτιμήτρια για την εκτίμηση όλου του αυθεντικού γραφήματος. Οι ακμές που είναι ορατές σε κάθε δείγμα ξέρουμε σίγουρα ότι ανήκουν και στο αυθεντικό γράφημα ενώ για τα ζεύγη ακμών που δεν έχουν ακμή στο δείγμα δεν ξέρουμε αν αυτό οφείλεται στο ότι η ακμή δεν υπάρχει στο αυθεντικό γράημα ή υπάρχει και απλά είναι κρυμμένη. Η πιο φυσιολογική εκτιμήτρια που θα μπορούσαμε να σκεφτούμε σε αυτό το σημείο είναι η ένωση όλων των δειγμάτων $G_1, \ldots, G_N$.

$$(\hat{G})_{ML} = \bigcup_{i=1}^{N} G_i$$

Μπορούμε να δείξουμε ότι αυτή η εκτιμήτρια είναι και εκτιμήτρια μέγιστης πιθανοφάνειας (ΕΜΠ), αφού πρέπει να συμπεριλάβουμε όλες τις ακμές της ένωσης για να μην μηδενιστεί η πιθανοφάνεια και επιπλέον, αν συμπεριλάβουμε περισσότερες, η πιθανοφάνεια δεν πρόκειται να αυξηθεί.

Τώρα προσδιορίζουμε την δειγματική πολυπλοκότητα αυτής της εκτιμήτριας. Πόσα δείγματα χρειάζονται για να ισχύει η (1); Είναι εύκολο να δούμε ότι με λογαριθμικά πολλά δείγματα $\log(n/\delta)$, όλες οι κορυφές θα γίνουν σε κάποιο από τα δείγματα ορατές, φανερώνοντας την γειτονιά τους και άρα όλο το γράφημα θα γίνει ορατό με πιθανότητα τουλάχιστον $1 - \delta$.

Πιο ενδιαφέρον είναι να εξετάσουμε αν αυτός είναι και απαραίτητος αριθμός δειγμάτων από κάθε εκτιμήτρια που πετυχαίνει την (1). Με επιχειρήματα από την θεωρία πληροφορίας, και συγκεκριμένα με χρήση της ανισότητας του Fano, απαντούμε ότι πράγματι λογαριθμικό πλήθος δειγμάτων είναι αναγκαίο.

Θεώρημα 1. *Υπάρχει μια εκτιμήτρια $\hat{G}$ τέτοια ώστε για κάθε παράμετρο $\delta \in (0,1]$ και γράφημα G, δεδομένων $N = \Theta(\log(n/\delta))$ ανεξάρτητων δειγμάτων $X_1, \ldots, X_N \sim \mathcal{P}_G$ ικανοποιεί $\mathbb{P}_{\mathbf{X} \sim \mathcal{P}_G^N}(\hat{G} \neq G) < \delta$. Επιπλέον, αν $N = o(\log n)$ τότε για κάθε εκτιμήτρια $\hat{G}$ υπάρχει γράφημα G τέτοιο ώστε $\mathbb{P}_{\mathbf{X} \sim \mathcal{P}_G^N}(\hat{G} \neq G) \geq 1/3$.*

Η ιδέα για την απόδειξη του παραπάνω θεωρήματος είναι ότι ορίζουμε μια οικογένεια γραφημάτων και δείχνουμε ότι οποιοσδήποτε αλγόριθμος που καταφέρνει να

ξεχωρίσει ποιο γράφημα της οικογένειας γέννησε τα δείγματα χρειάζεται τουλάχιστον $\log n$ δείγματα. Αυτό το σενάριο είναι γνωστό στην στατιστική ως *έλεγχος στατιστικών υποθέσεων*.

Προσεγγιστική Εκτίμηση Ακμών

Προχωρώντας στην προσεγγιστική εκτίμηση ακμών, αρχικά εξετάζουμε την εκτιμήτρια που απλά μετράει τις ακμές στο δείγμα G_s και τις κανονικοποιεί κατάλληλα.

$$\hat{m} = \frac{|E(G_s)|}{1-p^2} \tag{2}$$

Φαίνεται εύκολα ότι $\mathbb{E}[\hat{m}] = |E(G_s)|$. Για τον υπολογισμό της διασποράς αυτής της εκτιμήτριας, υπάρχουν όροι από τη διασπορά κάθε ακμής του τυχαίου δείγματος καθώς και όροι συνδιακύμανσης μεταξύ ακμών που μοιράζονται κοινή κορυφή. Αφού υπολογιστούν οι όροι αυτοί και ύστερα από μερικές πράξεις προκύπτει ότι

$$\mathrm{Var}(\hat{m}) \leq \frac{p^2}{2(1-p^2)} \sum_{v \in V} d_G^2(v)$$

Επειδή τελικά θέλουμε να εξασφαλίσουμε την (2), εφαρμόζουμε την ανισότητα Chebysev για να πάρουμε τον τύπο για το σφάλμα

$$\mathbb{P}(|\hat{m} - m| > \varepsilon m) \leq \frac{\mathrm{Var}(\hat{m})}{\varepsilon^2 m^2} \leq \frac{p^2}{2\varepsilon^2(1-p^2)} \frac{\sum_{v \in V} d_G^2(v)}{m^2}$$

Παρατηρούμε ότι εξαρτάται από τον λόγο $\frac{\sum_{v \in V} d_G^2(v)}{m^2}$ το αν θα είναι το σφάλμα μικρό ή μεγάλο. Για παράδειγμα, σε ένα d-κανονικό γράφημα ο λόγος αυτός πέφτει σαν $1/n$ και άρα η εκτίμηση βελτιώνεται με φυσικό τρόπο. Η χειρότερη περίπτωση είναι το γράφημα να είναι άστρο, όπου ο λόγος είναι σταθερά. Εκεί η κεντρική κορυφή έχει πολλές ακμές πάνω της και η συνεισφορά στη διασπορά είναι τεράστια. Παρακινούμενοι από αυτή την παρατήρηση, στην επόμενη ενότητα θα προσπαθήσουμε να βρούμε τέτοιες προβληματικές κορυφές και να ρωτάμε το μαντείο για την πραγματική γειτονιά τους ώστε να διορθώσουμε αυτή την εκτιμήτρια.

Πριν το κάνουμε αυτό, προσδιορίζουμε πάλι το πλήθος δειγμάτων που χρειάζεται για την εκτίμηση, για να το συγκρίνουμε με το πλήθος δειγμάτων της διορθωμένης εκτιμήτριας που χρησιμοποιεί το μαντείο (εκείνη προφανώς θα θέλουμε να δουλεύει με λιγότερα δείγματα). Η (2) όπως είναι γραμμένη χρησιμοποιεί ένα δείγμα, αλλά μπορούμε να την επεκτείνουμε να δουλεύει με N απλά χρησιμοποιώντας την ένωσή τους σαν ένα καλύτερο δείγμα που έχει παραχθεί από το ίδιο μοντέλο με παράμετρο $p' = p^N$.

Θεώρημα 2. *Υπάρχει μια εκτιμήτρια $\hat{m}$ τέτοια ώστε για κάθε $\varepsilon, \delta \in (0,1]$ και γράφημα G με m ακμές, δεδομένων $N = \Theta(\log(1/\varepsilon^2\delta))$ ανεξάρτητων δειγμάτων $X_1, \ldots, X_N \sim \mathcal{P}_G$ ικανοποιεί $\mathbb{P}_{\mathbf{X} \sim \mathcal{P}_G^N}(|\hat{m} - m| > \varepsilon m) < \delta$. Επιπλέον, αν $N = o(\log \varepsilon^{-1})$ τότε για κάθε εκτιμήτρια $\hat{m}$ υπάρχει ένα γράφημα G τέτοιο ώστε $\mathbb{P}_{\mathbf{X} \sim \mathcal{P}_G^N}(|\hat{m} - m| > \varepsilon m) \geq 1/3$.*

Για την απόδειξη του κάτω φράγματος σε αυτό το θεώρημα, γίνεται αναγωγή σε έλεγχο στατιστικών υποθέσεων, παρόμοιο με αυτόν του προηγούμενου θεωρήματος.

Προσεγγιστική Εκτίμηση Τριγώνων

Η εκτίμηση τριγώνων είναι παρόμοια με αυτή των ακμών με μόνη διαφορά ότι οι πράξεις γίνονται περισσότερες. Τελείως αντίστοιχα, συμβολίζοντας με $T(G)$ το σύνολο τριγώνων του γραφήματος G, το σημείο αφετηρίας είναι η εκτιμήτρια

$$\hat{T} = \frac{|T(G_s)|}{1 - 3p^2 + 2p^3}$$

όπου στον παρονομαστή έχουμε την πιθανότητα να είναι ένα τρίγωνο ορατό στο δείγμα, ώστε τελικά η αναμενόμενη τιμή της $\hat{T}$ να είναι ίση με $|T(G)|$. Ο υπολογισμός της διασποράς εδώ χρειάζεται πολλές πράξεις. Συμβολίζουμε με $\lambda(v)$ το πλήθος τριγώνων που ακουμπούν στην v και με $\lambda(u, v)$ το πλήθος τριγώνων που περιλαμβάνουν την ακμή $(u, v) \in E$. Τελικά βρίσκουμε

$$\mathbb{P}(|\hat{T} - T| > \varepsilon T) \leq \frac{\mathrm{Var}(\hat{T})}{\varepsilon^2 T^2} \leq \frac{c_1(p) \sum_{v \in V} \lambda^2(v) + c_2(p) \sum_{(u,v) \in E} \lambda^2(u, v)}{\varepsilon^2 T^2}$$

όπου $c_1(p) = 6p^3(1-p)^3/(1-3p^2+2p^3)^2$ και $c_2(p) = p^2(1-p)^2(1-2p)^2/(1-3p^2+2p^3)^2$ είναι σταθερές που εξαρτώνται μόνο από το p.

Η αντιστοιχία με την περίπτωση των ακμών είναι εμφανής. Αντί για βαθμούς, τώρα στον αριθμητή εμφανίζονται "βαθμοί τριγώνων". Οπότε, οι κορυφές που συνεισφέρουν περισσότερο στο σφάλμα είναι αυτές που πάνω τους ακουμπούν πολλά τρίγωνα. Η δειγματική πολυπλοκότητα για την μάθηση τριγώνων είναι ίδια με αυτή των ακμών και η απόδειξη είναι πανομοιότυπη.

Θεώρημα 3. *Υπάρχει μια εκτιμήτρια $\hat{t}$ τέτοια ώστε για κάθε $\varepsilon, \delta \in (0, 1]$ και γράφημα G με t τρίγωνα, δεδομένων $N = \Theta(\log(1/\varepsilon^2\delta))$ ανεξάρτητων δειγμάτων $X_1, \ldots, X_N \sim \mathcal{P}_G$ ικανοποιεί $\mathbb{P}_{\mathbf{X} \sim \mathcal{P}_G^N}(|\hat{t} - t| > \varepsilon t) < \delta$. Επιπλέον, αν $N = o(\log \varepsilon^{-1})$ τότε για κάθε εκτιμήτρια $\hat{t}$ υπάρχει ένα γράφημα G τέτοιο ώστε $\mathbb{P}_{\mathbf{X} \sim \mathcal{P}_G^N}(|\hat{t} - t| > \varepsilon t) \geq 1/3$.*

Εκτίμηση από Δείγματα και Χρήση Μαντείου

Όπως φάνηκε από τα προηγούμενα, η συνεισφορά κάθε κορυφής στο variance της εκτίμησης των ακμών είναι ανάλογη του βαθμού της στο τετράγωνο. Η ιδέα είναι ότι αν αρχικά το σφάλμα της εκτίμησης είναι μεγάλο, αυτό οφείλεται στο γεγονός ότι το αυθεντικό γράφημα μοιάζει σαν μια μικρή συλλογή από άστρα, δηλαδή υπάρχουν λίγες κορυφές που πάνω τους ακουμπούν σχεδόν όλες οι ακμές του γραφήματος. Αν ξέραμε ποιο είναι αυτό το σύνολο Q των κρίσιμων κορυφών, θα μπορούσαμε να ρωτήσουμε για την πραγματική γειτονιά τους και να αξιοποιήσουμε αυτή την πληροφορία για βελτιώσουμε την εκτιμήτρια. Πιο συγκεκριμένα θα χρησιμοποιούσαμε την εκτιμήτρια

$$\hat{m} = \sum_{e=(u,v) \in E} \mathbf{1}(u \in Q \vee v \in Q) + \frac{1}{1-p^2} \sum_{e \in E(G[V \setminus Q])} \mathbf{1}(e \in E(G_s))$$

Μπορούμε να δείξουμε ότι αν το Q περιέχει τις $\Theta(\varepsilon^{-2}\delta^{-1})$ κορυφές μέγιστου βαθμού, τότε το ε-σχετικό σφάλμα αυτής της εκτιμήτριας είναι πάντα κάτω από δ. Το ενδιαφέρον είναι ότι το πλήθος των κορυφών στο Q είναι σταθερό ως προς το μέγεθος

του γραφήματος. Βέβαια, μέχρι τώρα είπαμε τι θα κάναμε αν ξέραμε τις μεγιστοβάθμιες κορυφές του αυθεντικού γραφήματος, που προφανώς είναι κάτι μη εφικτό. Η ιδέα είναι ότι αν ορίσουμε το Q να περιέχει τις $\Theta(\varepsilon^{-2}\delta^{-1})$ μεγιστοβάθμιες κορυφές του ενός δείγματος, το ποσό που θα μειωθεί το σφάλμα θα είναι μια προσεγγιστικά ίδιο με την μείωση που θα παίρναμε αν χρησιμοποιούσαμε τις βέλτιστες κορυφές. Έτσι, χρησιμοποιούμε δύο δείγματα G_1, G_2, όπου από το G_1 καθορίζουμε το σύνολο $Q(G_1)$ των $\Theta(\varepsilon^{-2}\delta^{-1})$ μεγιστοβάθμιων κορυφών και η τελική εκτίμηση που υπολογίζεται με βάση το G_2 είναι η

$$\hat{m} = \sum_{(u,v)\in E(G)} \mathbf{1}(u \in Q(G_1) \vee v \in Q(G_1)) + \frac{1}{1-p^2} \sum_{e \in G[V\setminus Q]} \mathbf{1}(e \in E(G_2)) \qquad (3)$$

Η ανάλυση που μόλις περιγράψαμε πολύ περιληπτικά τελικά οδηγεί το παρακάτω συμπέρασμα (που δίνουμε μια ανεπίσημη εκδοχή του εδώ).

Θεώρημα 4. *Υπάρχει μια εκτιμήτρια $\hat{m}$ η οποία χρησιμοποιεί δύο δείγματα και $k = \Theta(\varepsilon^{-2}\delta^{-1})$ ερωτήματα προς το μαντείο και ικανοποιεί $\mathbb{P}(|\hat{m} - m| > \varepsilon m) \leq \delta$.*

Εκτίμηση Πλήθους Τριγώνων

Η ιδέα είναι παρόμοια με αυτή για τις ακμές και έτσι τα αποτελέσματα γενικεύονται και για τα τρίγωνα. Θα πρέπει να εκτιμήσουμε και εδώ από το δείγμα ποιες είναι οι κορυφές που πάνω τους ακουμπούν πολλά τρίγωνα στο αυθεντικό γράφημα και να τις ρωτήσουμε για να μειώσουμε την συνεισφορά τους στο σφάλμα. Η διαφορά σε σχέση με την περίπτωση των ακμών είναι ότι αυτό το εγχείρημα είναι δυσκολότερο. Πράγματι, οι βαθμοί κορυφών διατηρούνται στα δείγματα (οι βαθμοί είναι είτε ίδιοι είτε ακολουθούν διωνυμική κατανομή) ενώ το πλήθος τριγώνων που ακουμπά σε κάθε κορυφή όχι (φανταστείτε πολλά τρίγωνα που μοιράζονται μία κοινή ακμή, αν αυτή χαθεί χάνονται όλα τα τρίγωνα). Οπότε φαίνεται ότι δεν θα είναι εφικτό να πετύχουμε την εγγύηση (2). Παρόλλα αυτά αν χρησιμοποιήσουμε πολλαπλασιαστικό σφάλμα εW, όπου W ο αριθμός των *σφηνών*, δηλαδή τριάδων κορυφών με δύο ακμές (σαν τρίγωνα αλλά με μία ακμή να λείπει), αντί για σφάλμα εT, τότε τα αποτελέσματα γενικεύονται.

Θεώρημα 5. *(Ανεπίσημο) Υπάρχει εκτιμήτρια $\hat{T}$ που χρησιμοποιεί δύο δείγματα και $k = \Theta(\varepsilon^{-2}\delta^{-1})$ ερωτήματα προς το μαντείο και ικανοποιεί $\mathbb{P}(|\hat{T} - T| > \varepsilon W) \leq \delta$, όπου W είναι το συνολικό πλήθος σφηνών στο γράφημα G.*

Contents

Chapter 1

Introduction

The subject of this thesis is estimation of unknown parameters, which have always been a central problem in statistics. Algorithms can be divided into categories depending on their purpose. On the one hand, some algorithms are used to calculate functions, that is, output a specific result for each possible input. In this case, time complexity which is the amount of time it takes to run the algorithm is the main measure of efficiency. In this context, computer scientists begun (and have never stopped) to ask which are those functions that can be computed efficiently and which cannot. On the other hand, the purpose of some other algorithms is to estimate missing or unknown information, such as the expected value of a distribution by taking as input samples drawn from that distribution. The measure of efficiency in that case, apart from time complexity, is the sample complexity which is the number of samples required for an accurate estimation. Before we go into the details of the problem examined in this thesis, we briefly discuss its context.

Social networks

Graphs are ubiquitous structures for representing information and are used in many diverse fields, such as information systems, biology and social networks. In this thesis we focus mainly on social networks which are used to encode friendship relations between humans. Social networks are undirected graphs which have a vertex for each user and an edge for each pair of users that are friends. Much effort has been made by computer scientists and mathematicians to derive random graph models with the same properties as those of real world social networks, such as skewed degree distribution [New03] or small average diameter [AB02, B$^+$16, DM13]. Also, the size of these graphs is typically very large and can reach millions of vertices and edges. Thus, fast algorithms are required in order to run in feasible amounts of time.

Why Triangles

But what tasks do we want efficient algorithms for? Triangles have received great attention in the field of social network analysis as they are closely related to some significant structural properties of graphs. The creation of triangles in social networks is explained by certain laws. The first one is that people with common

friends tend to become friends themselves, a phenomenon known as *transitivity* and the second is that people who are friends statistically have similar interests and thus tend to have many common friends, a phenomenon called *homophily* [WF+94]. Therefore, in order for these characteristics of social networks to be quantified, two metrics have been defined and are extensively used in the literature. The *transitivity ratio* [NWS02] of a graph G is the probability that a wedge (which is defined as a path of length two), which is selected uniformly at random from all wedges of G, will participate in a triangle. The *clustering coefficient* [WS98] of a vertex v is defined as the probability that two uniformly selected neighbors of v will be connected with an edge. In addition, counting the number of triangles is used in a number of graph mining applications such as spam detection and community detection [BBCG08, EM02](for a longer list see [TKM11]). Therefore, new triangle counting algorithms are continuously proposed and studied from both a practical and a theoretical perspective.

Motivation and Statement of the Problem

There is a considerable amount of literature on triangle counting or other algorithms that can run fast on massive networks, however, in this thesis we examine the problem from a learning viewpoint. Privacy plays a major role in social networks, where each user can mark some of her friends as private, which makes them a perfect example of limited knowledge environments. In [CEK+15], this privacy setting is examined, where the *public-private* model of graphs is introduced and algorithms for it are presented. In that model, the graph G which is known to the public and does not contain edges between users that decided to hide their friends is called *public*, while a *private* graph G_u associated with each user u contains that user along with all her private friends. The authors were interested in designing algorithms that would preprocess the public graph and then would very efficiently compute properties of the graphs $G \cup G_u$, which essentially are the social network from the viewpoint of each user. Thus, these algorithms would run in a tailored way for each user, respecting the privacy constraints of other users. The algorithms presented were also efficient, meaning that they were much faster than running the same algorithm for the property of interest on the union $G_u \cup G$ one time for each user.

Motivated by the characteristic of social networks to have public and private vertices and also from [CEK+15], we go a step further and ask what can be done about estimating properties of the union of the public with all private graphs $G \cup \{G_u\}_u$, having access to only the public graph. Therefore, in this thesis, we treat the process of marking private friends as noise, which removes parts of the graph, and we are interested in recovering information about the initial graph. Based on this idea we are lead to propose a variant of the public-private model.

In our model we have the *underlying* graph G, which contains all connections between friends and each user decides whether to hide its friends with some probability p. If two friends make this decision their edge is removed from the resulting noisy sample graph G_s. Admitting that, based solely on the noisy sample graphs the learning tasks can be extremely hard, we allow our algorithms to have additional access to a small number of private graphs, equivalently, we allow them to query for

the true neighborhood of some vertices. In this model, we are interested in deriving estimators for interesting parameters of the underlying graph, such as the number of triangles.

Contribution

In this thesis, we begin by deriving and studying simple estimators, which only use samples of the graph, for the number of edges and triangles under the aforementioned model. More specifically, they output an estimation that is whithin a certain multiplicative factor away from the real value, with high probability. For these, we find that their sample complexity is optimal for the specific tasks by constructing an information theoretic lower bound.

Based on this examination, we then derive estimators for the same tasks that also perform a certain number of queries. We show that by allowing queries, the number of samples needed for accurate estimation drops to constant and also the number of queries is small, as it does not depend on the size of the graph.

Related Work

As mentioned before, due to the significance of triangles in social networks, in this thesis we focus on deriving estimators for their number in the noise model motivated by the privacy characteristics of social networks. Much work exists on triangle counting, ranging from exact counting algorithms, based on matrix multiplication [AYZ97, CW90] to approximate counting algorithms. In this section, we provide a quick overview of some of these ideas that are most relevant to this thesis.

Graph Sparsifiers

Approximation algorithms that use samples of the original graph were proposed for triangle counting in order to achieve faster runtimes. In these works, samples are used in a different context than in this thesis, were we treat samples as the result of noise altering the original graph. More specifically, samples are used as a mean of sparsification, that is, obtaining a subgraph of significantly smaller size from which the triangles of the original graph can be accurately estimated by counting their number in the sample and scaling them properly. Tsourakakis et al started these sparsification mechanisms, the most important of which is Doulion [TKMF09]. In Dulion, the sparsification process of G consists of keeping each edge with probability p and deleting it with $1-p$, resulting in a subgraph G'. After that, an exact counting algorithm is applied to count the number of triangles in G' and the result is scaled by $1/p^3$ to obtain the desired estimate X. This algorithm is very accurate in practice. In [TKM11, TKM09] a more thorough analysis is done to derive strong theoretical guarantees using some advanced concentration inequalities for random variables that are polynomials.

Graph sampling can be thought of as a basic preprocessing step of the input before the computation of any graph property. The goal always remains the same, namely to produce a smaller graph that in a way preserves the desired property of the original graph, making its computation more efficient. Some sampling methods are

edge sampling, such as Doulion, vertex sampling [LKJ06], where one choses random sets of vertices and examines the corresponding induced subgraph and traversal based sampling [LF06], which can be a random walk on the graph's vertices. For a complete survey on the topic we refer to [HL13].

Learning via Queries

In the previous type of works, the whole graph has to be read for its sparsification. However, another line of research is dedicated on algorithms that are truly sublinear (meaning that they do not need to read the entire graph) in order to achieve even lower time complexities. We would like to think of these algorithms as having restricted knowledge of the input, which is more on par with the context of this thesis. In other words, these are essentially learning algorithms that need to extract properties of an unknown graph. *Sublinear algorithms* is a growing field that initialy started with property testing [Gol17], the task to decide whether an object has a property or is far away from having it, but recently has embraced estimation problems too. Access to graph G is available by queries of three types that are now standard in the literature:

1. Degree query: The degree $d(v)$ of any vertex $v \in V$.
2. Neighbor query: What vertex is the i-th neighbor of vertex $v \in V$ (where some arbitrary ordering of the neighbors is assumed and also $i < d(v)$).
3. Pair query: Test whether $(u, v) \in E$ for any pair of vertices $(u, v) \in V^2$.

The naive way to count triangles in this setting is by a simple Monte Carlo algorithm: sampling random triplets of vertices and checking if they form a triangle in the graph. By executing k trials and counting the number of successes (when triplets formed triangles) it is a simple exercise of concentration inequalities to see that if k is big enough we have a good approximation.

The first non trivial algorithm for triangle estimation was that of [ELRS15]. For a graph with n vertices and m edges, an estimator $\hat{t}$ for the number of triangles was designed that works with $O\left(\varepsilon^{-1}(\text{poly}\log n)\left(\frac{n}{t^{1/3}} + \frac{m^{3/2}}{t}\right)\right)$ standard type queries and satisfies $\mathbb{P}(|\hat{t} - t| > \varepsilon t) < 1/3$. The algorithm's analysis is quite intricate and requires a number of ideas. In [Ses15], an algorithm with the same guarantees but simpler analysis was proposed.

In this setting, other properties were examined too before triangles, such as estimating the number of stars [GRS11] or cliques of fixed size [ERS18]. After the triangle estimation algorithm, the result was generalized for arbitrary subgraph counting [AKK18].

The relevance of these works to this thesis is that we would like to combine elements from both the sparsification methods mentioned before and the query based algorithms in our work. We note that our goal is to calculate an initial rough estimation of the number of triangles from a single noisy sample graph, which is very similar to what algorithms like Doulion do, and then, given additional query access, improve upon this estimation by asking vertices that are critical to the number of triangles. This is something that, to the best of our knowledge, has not been studied in the literature yet.

Online Learning for Network Discovery

Apart from the fact that the graph inputs are too large to be handled, which leads to the need for sparcification, the input may be incomplete by its self. For example graph data for social networks are collected from apps or users who make their accounts public and are inherently incomplete. Thus, a research direction focuses in enhancing these partially observed graphs via probing parts of the network. A model for the queries is each vertex can be asked to reveal all of its neighbors or just one uniformly at random. The question is, given a limited budget of such queries, how can we efficiently probe the graph so as to reveal the greatest number of new vertices or triangles or, in its full generality, optimize an arbitrary graph function? The model describing how the incomplete input is generated does not need to be known and the task becomes an online learning problem [LSBER18]. The problem has also been formulated as a multi-armed bandit problem and an algorithm that includes both exploration and exploitation proccedures with efficient performance in practice has been proposed [SERGP17].

Thesis Organization

This is a brief outline of the content of each chapter to guide the reader through this thesis. Effort has been made to keep chapters concise enough to preserve readability and yet complete and self contained, regarding the way the ideas of this thesis are analyzed. All preliminaries are located in Chapter 2. More specifically, in that chapter, basic concentration inequalities that will be used later are presented and well known tools from hypothesis testing and information theory that are usefull for lower bound construction are developed.

Chapter 3 is about parameter estimating using only noisy samples of the original graph. First, the noise model is formally presented and then estimators for the number of edges and triangles as well as for the whole graph are given. Their sample complexity is determined and lower bounds using the tools of Chapter 2 are constructed to understand their limitations.

In Chapter 4, queries are allowed to improve the accuracy of estimation, or equivalently, reduce the number of samples required. Based on the conclusions from the analysis of the previous chapters, we develop a query strategy and design an estimator for edges and triangles that embraces information from these queries.

Finally, Chapter 5 contains our conclusions and directions for ongoing or future work on the subject.

Chapter 2

Information Theoretic Lower Bounds

A part of our work will be focused on constructing lower bounds on the number of samples needed to learn a property of the underlying distribution. Proving that no algorithm can achieve better performance for a specific task is essential for understanding a problem and is often more interesting than designing an algorithm for it. Lower bounds can be challenging to construct, yet mathematicians' efforts have resulted in many frameworks for that purpose. In order to prove that an estimator needs at least a certain number of samples, we need to argue that less samples do not provide enough content for recognizing which source could have generated the samples. The concept of *useful content* is captured by the tools of information theory. In this chapter we provide an overview of the tools and machinery that are necessary to construct lower bounds. For the shake of completeness we start by some preliminaries about concentration of measure that, although not exploited during the lower bound derivation, are used in later chapters.

2.1 Statistics Preliminaries

Often in statistics and in the analysis of randomized algorithms we need to use the fact that random variables are close to their expected value with high probability. The motivation can become clear with a simple example. Consider tossing a fair coin n times. After many trials, that is, if n is great enough, half of the results will be heads with high confidence. We need to quantify this behavior. The law of large numbers states that the mean of many random variables converges almost surely to the expected value, as their number increases. However, the law of large numbers or other results such as the central limit theorem are asymptotic results. We would like to determine exact bounds on the rate of convergence, which is what the following theorems essentially do.

Theorem 2.1 (Markov's Inequality)**.** *Let X be a non negative random variable. For any $a > 0$*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof.

$$\mathbb{E}[X] = \sum_x x\,\mathbb{P}(X = x) = \sum_{x \geq a} x\,\mathbb{P}(X = x) + \sum_{x < a} x\,\mathbb{P}(X = x) \geq a\,\mathbb{P}(X \geq a)$$

The fact that X must be non negative was required in order to bound the second sum. ■

The next inequality is derived from Markov's inequality and bounds from above the probability of a random variable to deviate from its expected value.

Theorem 2.2 (Chebysev's Inequality)**.** *Let X be a random variable. Then for any $a > 0$*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2} \tag{2.1}$$

Proof. The result follows from an application of Markov's inequality to the random variable $(X - \mathbb{E}[X])^2$ and the observation that $\mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{Var}[X]$. ■

Next, we present Chernoff bounds that are used for sums of random variables. There exist many variants of these theorems, with the more general of them having information theoretic essence, as they are expressed in therms of the mutual information of the distributions involved. For a complete exposition on the subject see [BLM13]. Here we only state the variant we will use later in our analyses which is the multiplicative bound.

Theorem 2.3 (Chernoff Bound)**.** *Let $X_1, X_2, \ldots, X_n$ be independent $\{0, 1\}$ random variables with $\mathbb{E}[X_i] = p_i$ and define $X = \sum_{i=1}^{n} X_i$. Let $\mu = \mathbb{E}[X]$ denote the expected value of X. For every $0 < \varepsilon < 1$ it holds*

$$\mathbb{P}(X > (1 + \varepsilon)\mu) \leq \exp\left(-\frac{\varepsilon^2 \mu}{3}\right) \tag{2.2}$$

$$\mathbb{P}(X < (1 - \varepsilon)\mu) \leq \exp\left(-\frac{\varepsilon^2 \mu}{2}\right) \tag{2.3}$$

Proof. For the standard proof that uses the moment generating function we refer the reader to page 66 of [MU17]. ■

These bounds are preferred over the Chebysev's inequality, when applicable, because they demonstrate a much more intense concentration of measure. Indeed, these bounds are tight within a constant factor [Mou10].

2.2 Divergence Measures

Divergence measures are functions that establish the distance between probability distributions. A very formal definition of these concepts, which can be found in [Gra11], requires a bit of setup and also understanding of measure theory. Here we present the definitions only to an extend that serves our purposes.

Let $\mathcal{X}$ be an arbitrary space, $\mathcal{P}, \mathcal{Q}$ two distributions on that space and also let p, q denote their probability density functions. For every $A \subset \mathcal{X}$, each distribution assigns a probability which will be denoted by $P(A), Q(A)$.

Definition 2.1. *The total variation distance between $\mathcal{P}$ and $\mathcal{Q}$ is defined as*

$$\|\mathcal{P} - \mathcal{Q}\|_{TV} \triangleq \sup_{A \subset \mathcal{X}} |P(A) - Q(A)| = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| dx$$

Thus, if we imagine the two distributions as hills of sand, their total variation distance can be thought of as the amount of sand that must be moved from one to another to make them equal. For discrete random variables, integrals are replaced by sums. Next, we define another divergence which technically is not a distance. This divergence is a member of a well known family of divergences called f-divergences [AS66, Csi67].

Definition 2.2. *The Kullback-Leibler divergence between $\mathcal{P}$ and $\mathcal{Q}$ is*

$$D_{\mathrm{KL}}(\mathcal{P} || \mathcal{Q}) \triangleq \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx$$

We finish this section with two properties of KL-divergence. The first one, which is easier to prove, is about products of distributions $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n, \mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_n$. For these, KL-divergence satisfies

$$D_{\mathrm{KL}}(\mathcal{P} || \mathcal{Q}) = \sum_{i=1}^{n} D_{\mathrm{KL}}(\mathcal{P}_i || \mathcal{Q}_i) \tag{2.4}$$

The second property is Pinsker's inequality, a result that required more effort by mathematicians in order to be proved. For arbitrary distributions $\mathcal{P}$ and $\mathcal{Q}$ (not only products as we assumed before) we have

$$\|\mathcal{P} - \mathcal{Q}\|_{TV}^2 \leq \frac{1}{2} D_{\mathrm{KL}}(\mathcal{P} || \mathcal{Q}) \tag{2.5}$$

2.3 Information Theory Concepts

Here we present the basic definitions and properties of information theoretic concepts that will be used later in this thesis. This is not a complete investigation but a rather quick review of the elements that our work will depend on. For further information we refer the reader to [CT12].

2.3.1 Entropy

For the following we will let capital letters such as X denote random variables and the corresponding calligraphic letters such as $\mathcal{X}$ denote the alphabets from which they take values.

Definition 2.3. *The entropy of a random variable X with probability mass function $p(x)$ is defined as*

$$H(X) \triangleq -\sum_{x \in \mathcal{X}} p(x) \log p(x) = \mathbb{E}[-\log p(X)]$$

We will adopt the convention $0 \log 0 = 0$ because of the limit behavior of the function $x \log x$ when x goes to zero. Entropy serves as a measure of the information content of a random variable. Consider for example a degenerated random variable that is just a constant. Its entropy is zero. Also, $H(X) \geq 0$ for every random variable X and more importantly

Proposition 2.1. *$H(X) \leq \log |\mathcal{X}|$ with the inequality being tight in the case of a uniform random variable.*

The proof is deferred to a later subsection. Next we introduce joint and conditional entropy.

Definition 2.4. *The joint entropy of random variables X and Y is defined as*

$$H(X, Y) \triangleq -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

Definition 2.5. *The conditional entropy of Y given X is defined as*

$$\begin{aligned} H(Y \mid X) &\triangleq -\sum_{x \in \mathcal{X}} p(x) H(Y \mid X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \end{aligned}$$

An easy to prove property is the *chain rule*, which is

$$H(X, Y) = H(Y) + H(X \mid Y) \tag{2.6}$$

2.3.2 Mutual Information

Mutual information between two random variables, say X and Y, quantifies how much extra information about X is revealed when Y becomes known.

Definition 2.6. *The mutual information between random variables X and Y with joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$ is defined as*

$$\begin{aligned} I(X; Y) &\triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \left(\frac{p(x, y)}{p(x) p(y)} \right) \\ &= D_{\mathrm{KL}} \left(\mathcal{P}_{XY} || \, \mathcal{P}_X \mathcal{P}_Y \right) \end{aligned}$$

The mutual information can be rewritten in the following way (this is often given as the definition of mutual information). The proof is very easy and thus omitted.

Proposition 2.2. *For random variables X and Y we have $I(X; Y) = H(X) - H(X \mid Y)$.*

We can define conditional mutual information $I(X; Y|Z)$ similarly to conditional entropy, and we will have that $I(X; Y|Z) = H(X \mid Z) - H(X \mid Y, Z)$.

2.3.3 Information Inequality

We will prove the very important inequality of Proposition 2.1, which essentially tells us that the number of bits that are necessary to describe a random variable X are at most $\log|\mathcal{X}|$. First we begin by proving the following.

Proposition 2.3. *For random variables $X \sim \mathcal{P}_X$ and $Y \sim \mathcal{P}_Y$, it is true that $D_{\mathrm{KL}}(\mathcal{P}_X||\mathcal{P}_Y) \geq 0$ with the relation becoming an equality only if $\mathcal{P}_X = \mathcal{P}_Y$.*

Proof. Let $p(x), q(x)$ be the probability mass functions of X and Y respectively and $A = \{x \in \mathcal{X} \mid p(x) > 0\}$ the support of X.

$$\begin{aligned} -D_{\mathrm{KL}}(\mathcal{P}_X||\mathcal{P}_Y) &= -\sum_{x\in A} p(x)\log\left(\frac{p(x)}{q(x)}\right) \\ &= \sum_{x\in A} p(x)\log\left(\frac{q(x)}{p(x)}\right) \\ &\leq \log\sum_{x\in A} p(x)\frac{q(x)}{p(x)} \\ &= \log 1 \\ &= 0 \end{aligned}$$

where the inequality which was used was Jensen's inequality for concave functions. As $\log(\cdot)$ is strictly concave, the equality occurs iff $q(x)/p(x)$ is constant (independent of x), that is, $p(x) = q(x)$ for every x. ■

Corollary 2.1. *For random variables X and Y, we have that $I(X;Y) \geq 0$.*

Proof. The proof follows from noting that $I(X;Y) = D_{\mathrm{KL}}(\mathcal{P}_{XY}||\mathcal{P}_X\mathcal{P}_Y) \geq 0$ ■

Now we can prove the desired inequality.

Proof of Proposition 2.1. Let $p(x)$ be the probability mass function of X and $u(x)$ the probability mass function of the uniform distribution over alphabet $\mathcal{X}$.

$$u(x) = \begin{cases} \frac{1}{|\mathcal{X}|}, & x \in \mathcal{X} \\ 0, & x \neq \mathcal{X} \end{cases}$$

Let $\mathcal{P}$ be the distribution of X and $\mathcal{U}$ the uniform distribution. We have that

$$D_{\mathrm{KL}}(\mathcal{P}||\mathcal{U}) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{u(x)} = \log|\mathcal{X}| - H(X) \geq 0$$

because of the last corollary. ■

The following states that conditioning reduces entropy.

Corollary 2.2. *For random variables X and Y it is true that $H(X \mid Y) \leq H(X)$.*

Proof. $I(X;Y) = H(X) - H(X \mid Y)$ from Corollary 2.1. ■

2.3.4 Data Processing Inequality

The random variables X, Y, Z form a Markov chain $X \to Y \to Z$ if the conditional distribution of Z depends only on Y and not on X, that is, $p(z|x, y) = p(z|y)$. This means that given Y, the random variables X and Z become independent. The Markov chain presented may be interpreted in the following way. A process converts X to Y and another process is applied to Y and gives Z. It is natural to think that Z is less related to X than Y. In other words, no process (deterministic of randomized) can increase the information content about X. This is formally expressed by the following inequality.

Proposition 2.4 (Data Processing Inequality)**.** *Let X, Y, Z be random variables that form a Markov chain $X \to Y \to Z$. Then*

$$I(X;Y) \geq I(X;Z) \tag{2.7}$$

Proof.

$$\begin{aligned} I(X;Y,Z) &= H(X) - H(X \mid Y,Z) \\ &= H(X) - H(X \mid Z) + H(X \mid Z) - H(X \mid Y,Z) \\ &= I(X;Z) + I(X;Y|Z) \end{aligned} \tag{2.8}$$

We obtain the following in a similar manner.

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y) \tag{2.9}$$

By the fact that X, Z are independent the fact that given Y we have $I(X;Z|Y) = 0$. From (2.8) and (2.9) we get

$$\begin{aligned} I(X;Z) + I(X;Y|Z) &= I(X;Y) \\ I(X;Y) - I(X;Z) &= I(X;Y|Z) \geq 0 \\ I(X;Y) &\geq I(Y;Z) \end{aligned}$$

which is the desired relation. ■

2.4 Hypothesis Testing

We introduce the notion of testing statistical hypotheses, which was of central interest in the field of statistics [LR06]. The reason is that often we are given samples from a distribution for which we know that it belongs to a predefined set of distributions and we want to find which one of these was actually used to generate the samples. A *statistical hypothesis* is every hypothesis regarding the distribution of a random variable X. More specifically, a hypothesis H may regard a parameter $\boldsymbol{\theta}$ of the density function $p(x|\boldsymbol{\theta})$ or the form of the density function.

Testing hypotheses can be seen as a primitive form of estimation, where one does not need to output an exact value of the unknown parameter but just needs to distinguish between cases for that parameter. Although weaker than estimation, hypothesis testing can be as hard as estimation and for that reason it is very often

used to establish lower bounds for estimation problems, as we will show in the next section. We follow the exposition of [Duc16].

We define the setting of *canonical hypothesis testing*. Let $\mathcal{V}$ be a set of indices and $\{\mathcal{P}_v\}_{v\in\mathcal{V}}$ a family of distributions with support $\mathcal{X}$.

1. Nature chooses an index $V \in \mathcal{V}$ uniformly at random.
2. Conditioned on the event $V = v$, N samples are drawn from the v-th distribution $\mathbf{X} = (X_1, \ldots, X_N) \sim \mathcal{P}_v^N$.

The goal is to find the value v having access only to the samples $\mathbf{X}$. Every function $\Psi : \mathcal{X} \to \mathcal{V}$ used for that purpose is called a *test*. The joint distribution from which the samples $\mathbf{X}$ are drawn eventually is

$$\overline{\mathcal{P}} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathcal{P}_v^N$$

We will call a hypothesis testing problem *binary* if $|\mathcal{V}| = 2$, that is if there exist only two hypotheses. These hypotheses are denoted by H_0 and H_1, with H_0 being called *null hypothesis* and H_1 being called *alternative hypothesis*. If we have more than two hypotheses, the setting is referred as *multiple hypothesis testing*. A hypothesis may be *simple* or *composite*. The latter type of hypothesis is associated with a family of distributions instead of a single one. Conditioned on a composite hypothesis, nature chooses a distribution from that set uniformly at random and uses it to generate the samples. The testing problem is called *composite* if at least one of the hypotheses is of that type, otherwise it is called *simple*.

2.5 From Estimation to Hypothesis Testing

Providing solution to an estimation problem consists of designing an estimator that approximates the unknown value of interest well enough, meaning that some kind of guarantee is presented. For example, the empirical mean of N samples of Bernoulli variables is an unbiased estimator with variance that behaves like $1/N$. In order to completely solve the estimation problem, lower bounds have to be constructed, indicating that *every* estimator cannot do better than a certain threshold. Even for the simple example mentioned above, it is not obvious how to do this. This problem was resolved for the first time using the concept of *Fisher information* and the Cramer-Rao inequality [Cra46, Rao92]. A different way is described in this section, which is to first note that by having an estimator one could use it to solve a hypothesis testing regarding the unknown parameter, and then prove that this hypothesis testing problem is hard.

2.5.1 The Framework

To begin, we define the notion of the estimator's risk. Let $\mathcal{P}$ a distribution and $\theta(\mathcal{P})$ its parameter. A good estimator $\hat{\theta}$ for θ should have low $\mathbb{E}_{\mathbf{X}\sim\mathcal{P}^N}[(\hat{\theta}(\mathbf{X})-\theta(\mathcal{P}))^2]$. However, this definition of the risk would not be enough, as the estimator $\hat{\theta}(\mathbf{X}) = \theta(\mathcal{P})$ for every $\mathbf{X}$ would have zero risk, yet this estimator would be completely unreliable for other distributions than $\mathcal{P}$. This issue is resolved by selecting the

distribution used adversarially from the family of possible distributions $\mathcal{S}$, thus the *risk* is defined as the quantity

$$\sup_{\mathcal{P}\in\mathcal{S}} \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{P}^N} \left[(\hat{\theta}(\mathbf{X}) - \theta(\mathcal{P}))^2\right]$$

To establish a lower bound for every estimator, one needs to use the minimum value of this risk, where the min is taken over all estimators. This definition was introduced by Wald [Wal39].

Definition 2.7 (Minimax Risk)**.**

$$\mathfrak{M}(\theta, \mathcal{S}) \triangleq \inf_{\hat{\theta}} \sup_{\mathcal{P}\in\mathcal{S}} \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{P}^N} \left[(\hat{\theta}(\mathbf{X}) - \theta(\mathcal{P}))^2\right]$$

After having defined the risk, we focus on reducing the estimation problem to the hypothesis testing problem and lower bounding the former by providing a bound for the latter. Depending on whether the resulting hypothesis testing is binary or not, different inequalities are used to bound its error, resulting in different methods of constructing lower bounds. However, the reduction is always the same.

Let $\mathcal{V}$ be a set of indices, $\{\mathcal{P}_v\}_{v\in\mathcal{V}}$ a class of distributions indexed by $\mathcal{V}$ and $\{\theta(\mathcal{P}_v)\}_{v\in\mathcal{V}}$ their parameters.

Definition 2.8. *The class $\{\theta(\mathcal{P}_v)\}_{v\in\mathcal{V}}$ is called 2δ-packing if*

$$|\theta(\mathcal{P}_v) - \theta(\mathcal{P}_{v'})| \geq 2\delta \quad \forall v \neq v'$$

The following bounds the minimax risk of parameter estimation by the minimum error of a hypothesis testing problem.

Proposition 2.5. *Let $\mathcal{V}$ be a set of indices, $\mathcal{S} = \{\mathcal{P}_v\}_{v\in\mathcal{V}}$ a class of distributions indexed by $\mathcal{V}$ and $\{\theta(\mathcal{P}_v)\}_{v\in\mathcal{V}}$ their parameters such that $\{\theta(\mathcal{P}_v)\}_{v\in\mathcal{V}}$ is 2δ-packing. The minimax risk is bounded as follows*

$$\mathfrak{M}(\theta, \mathcal{S}) \geq \delta^2 \inf_{\Psi} \mathop{\mathbb{P}}_{\mathbf{X}\sim\overline{\mathcal{P}}}(\Psi(\mathbf{X}) \neq V)$$

Proof. The idea consists of the following two steps:

1. In order to solve the hypothesis testing regarding the family $\{\mathcal{P}_v\}_{v\in\mathcal{V}}$ we use the value of the estimator $\hat{\theta}(\mathbf{X})$ and return the distribution for which the parameter θ is closer to $\hat{\theta}(\mathbf{X})$. That is,

$$\Psi(\mathbf{X}) \triangleq \arg\min_{v\in\mathcal{V}} |\hat{\theta}(\mathbf{X}) - \theta(\mathcal{P}_v)|$$

2. We show that for that particular testing function Ψ, the probability of error is less than or equal to the minimax error of the estimation, that is

$$\mathfrak{M}(\theta, \mathcal{S}) \geq \delta^2 \mathop{\mathbb{P}}_{\mathbf{X}\sim\overline{\mathcal{P}}}(\Psi(\mathbf{X}) \neq V)$$

   However, Ψ is not the only testing function for the hypothesis testing and thus, the above expression can be further bounded by the infimum over all testing functions, resulting in the desired bound.

It remains to prove the argument of the second step. To do so, first we need to see when the testing function makes a wrong guess. Observe that because we have a 2δ-packing, if the testing function guesses wrongly, $\hat{\theta}$ and θ must be more than δ-away. To see that, suppose that $|\hat{\theta} - \theta(\mathcal{P}_v)| < \delta$. Then, for each $v' \neq v$, we have the inequality

$$2\delta \leq |\theta(\mathcal{P}_v) - \theta(\mathcal{P}_{v'})| \leq |\hat{\theta} - \theta(\mathcal{P}_v)| + |\hat{\theta} - \theta(\mathcal{P}_{v'})| < \delta + |\hat{\theta} - \theta(\mathcal{P}_{v'})|$$

which gives $|\hat{\theta} - \theta(\mathcal{P}_{v'})| > \delta$ and thus the test Ψ guesses correctly (a contradiction).

Next, we examine the risk

$$\begin{aligned}
\sup_{v\in\mathcal{V}} \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{P}_v^N} \left[(\hat{\theta}(\mathbf{X}) - \theta(\mathcal{P}_v)^2)\right] &\geq \sup_{v\in\mathcal{V}} \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{P}_v^N} \left[\delta^2 \mathbf{1}\left(|\hat{\theta}(\mathbf{X}) - \theta(\mathcal{P}_v)| \geq \delta\right)\right] \\
&\geq \delta^2 \sup_{v\in\mathcal{V}} \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}_v^N} \left(|\hat{\theta}(\mathbf{X}) - \theta(\mathcal{P}_v)| \geq \delta\right) \\
&\geq \delta^2 \frac{1}{|\mathcal{V}|} \sum_{v\in\mathcal{V}} \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}_v^N} (\Psi(\mathbf{X}) \neq v) \\
&= \delta^2 \mathop{\mathbb{P}}_{\mathbf{X}\sim\overline{\mathcal{P}}} (\Psi(\mathbf{X}) \neq V)
\end{aligned}$$

where for the third inequality we used the fact that the supremum of the elements of a set is greater than their mean value. Taking infimum over the possible estimators $\hat{\theta}$ finishes the proof. ■

What remains to do now is to present how the error of the hypothesis testing $\mathbb{P}_{\mathbf{X}\sim\overline{\mathcal{P}}}(\Psi(\mathbf{X}) \neq V)$ can be bounded from below.

2.5.2 Le Cam's Method

Le Cam's inequality [Yu97, LC12] is used in the case of binary hypothesis testing. The index set is now $\mathcal{V} = \{1, 2\}$ and the probability of error is

$$\mathop{\mathbb{P}}_{\mathbf{X}\sim\overline{\mathcal{P}}} (\Psi(\mathbf{X}) \neq V) = \frac{1}{2} \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}_1^N} (\Psi(\mathbf{X}) \neq 1) + \frac{1}{2} \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}_2^N} (\Psi(\mathbf{X}) \neq 2)$$

Proposition 2.6. *The error of the hypothesis testing is related to the total variation distance between the two distributions as follows*

$$\inf_{\Psi} \left\{ \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}_1^N} (\Psi(\mathbf{X}) \neq 1) + \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}_2^N} (\Psi(\mathbf{X}) \neq 2) \right\} = 1 - \|\mathcal{P}_1^N - \mathcal{P}_2^N\|_{TV} \tag{2.10}$$

Proof. Let A be the critical region of the testing function Ψ, that is the region where it assigns the input to the first distribution $\Psi(\mathbf{X}) = 1$. Also let $P_1(A), P_2(A)$ the probabilities assigned to region A by the N-fold distributions $\mathcal{P}_1^N, \mathcal{P}_2^N$. Then we have that

$$\mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}_1^N} (\Psi(\mathbf{X}) \neq 1) + \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}_2^N} (\Psi(\mathbf{X}) \neq 2) = P_1(A^c) + P_2(A) = 1 - P_1(A) + P_2(A)$$

Therefore, taking infimum over all testing functions gives

$$\begin{aligned}\inf_{\Psi}\left\{\mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}_1^N}(\Psi(\mathbf{X})\neq 1)+\mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}_2^N}(\Psi(\mathbf{X})\neq 2)\right\} &= \inf_{\Psi}\{1-P_1(A)+P_2(A)\}\\ &= 1-\sup\Psi\{P_1(A)-P_2(A)\}\\ &= 1-\|\mathcal{P}_1^N-\mathcal{P}_2^N\|_{TV}\end{aligned}$$

■

Combining this simple observation with the reduction of Proposition 2.5 we get a generic lower bound for the risk of estimation.

Corollary 2.3. *Let $\mathcal{P}_1,\mathcal{P}_2$ with $|\theta(\mathcal{P}_1)-\theta(\mathcal{P}_2)|\geq 2\delta$. The following is true for the minimax risk of estimation using N samples*

$$\mathfrak{M}(\theta,\mathcal{S})\geq\frac{1}{2}\delta^2\left(1-\|\mathcal{P}_1^N-\mathcal{P}_2^N\|_{TV}\right) \tag{2.11}$$

An observation about this bound is that there exists a trade off regarding the selection of δ. On the one hand, a lower δ means that the distributions have more similar value θ and thus are less distinguishable, something that is quantified by having small total variation distance. On the other hand, a smaller δ weakens the bound. Thus, the selection of δ should be the result of a fine tuning, in order to achieve the best value for the bound. We demonstrate the method with an example about Bernoulli variables.

Proposition 2.7. *The minimax lower bound for the estimation of the mean of a $\{-1,+1\}$ Bernoulli random variable using N samples is*

$$\mathfrak{M}\geq\frac{1}{24N}$$

Proof. We need to define two Bernoulli distributions $\mathcal{P}_1,\mathcal{P}_2$ with probabilities of success p_1,p_2 such that we have 2δ separation of them, as required by Corollary 2.3. We set

$$p_1=\frac{1+\delta}{2},\quad p_2=\frac{1-\delta}{2}$$

We have that

$$\begin{aligned}\|\mathcal{P}_1^N-\mathcal{P}_2^N\|_{TV}^2 &\leq \frac{1}{2}D_{\mathrm{KL}}\left(\mathcal{P}_1^N||\,\mathcal{P}_2^N\right)\\ &= \frac{N}{2}D_{\mathrm{KL}}\left(\mathcal{P}_1||\,\mathcal{P}_2\right)\\ &= \frac{N}{2}\left[\frac{1-\delta}{2}\log\left(\frac{1-\delta}{1+\delta}\right)+\frac{1+\delta}{2}\log\left(\frac{1+\delta}{1-\delta}\right)\right]\\ &= N\frac{\delta}{2}\log\left(\frac{1+\delta}{1-\delta}\right)\end{aligned}$$

where the first inequality is Pinsker's (2.5), and the second equality is property (2.4). To further bound the term we note that for $\delta < 1/2$ we have that

$$\delta \log\left(\frac{1+\delta}{1-\delta}\right) \leq 3\delta^2$$

To see that, set the corresponding function and differentiate it

$$\begin{aligned} f(x) &= \log\left(\frac{1+x}{1-x}\right) - 3x \\ f'(x) &= \frac{1-x}{1+x}\frac{(1-x)+(1+x)}{(1-x)^2} - 3 \\ &= \frac{2}{1-x^2} - 3 \end{aligned}$$

Setting $f'(x) = 0$ gives $x = \pm 1/\sqrt{3}$. For $x = 1/\sqrt{3}$ we have $f(1/\sqrt{3}) < 0$ and for $x = 0$ we have $f(x) = 0$, therefore $f(x) < 0$ in this region. Therefore, we get

$$\|\mathcal{P}_1^N - \mathcal{P}_2^N\|_{TV} \leq \delta\sqrt{\frac{3N}{2}}$$

By setting $\delta = 1/\sqrt{6N}$, (2.11) turns into

$$\begin{aligned} \mathfrak{M}(\theta, \mathcal{S}) &\geq \frac{1}{2}\delta^2\left(1 - \|\mathcal{P}_1^N - \mathcal{P}_2^N\|_{TV}\right) \\ &\geq \frac{1}{12N}\left(1 - \frac{1}{2}\right) \\ &= \frac{1}{24N} \end{aligned}$$

which is the bound that we wanted to prove. ■

2.5.3 Fano's Method

In this section we examine the inequality of Fano which bounds the error of multiple hypothesis testing. We will turn our attention to a slightly more general setting of moving information through a noisy channel, which includes hypothesis testing as a special case. Consider a Markov chain $X \to Y \to \hat{X}$. The interpretation is that Y is generated according to a random variable X (the domain of which will be denoted by $|\mathcal{X}|$), and then a process $\hat{X}$ which only has access to Y tries to recover X. Let $h_2(p) = -p\log p - (1-p)\log(1-p)$ denote the entropy of a Bernoulli random variable. We now state the inequality and provide its proof that is standard in the literature (see [CT12, Duc16] and also [Yu97] for some useful variants that are not presented here).

Proposition 2.8 (Fano's Inequality)**.** *Let $X \to Y \to \hat{X}$ be a Markov chain. It holds true that*

$$h_2(\mathbb{P}(\hat{X} \neq X)) + \mathbb{P}(\hat{X} \neq X)\log(|\mathcal{X}| - 1) \geq H(X \mid \hat{X})$$

Proof. We define a binary random variable E to be the indicator of the event $\hat{X} \neq X$.

$$E = \begin{cases} 1, & \hat{X} \neq X \\ 0, & \hat{X} = X \end{cases}$$

We expand the entropy $H(X, E \mid \hat{X})$ in the following two ways. The first one is

$$\begin{aligned} H(X, E \mid \hat{X}) &= H(X \mid E, \hat{X}) + H(E \mid \hat{X}) \\ &= \mathbb{P}(E = 1) H(X \mid E = 1, \hat{X}) + \mathbb{P}(E = 0) H(X \mid E = 0, \hat{X}) + H(E \mid \hat{X}) \\ &= \mathbb{P}(E = 1) H(X \mid E = 1, \hat{X}) + H(E \mid \hat{X}) \end{aligned}$$

where we used the fact that $H(X \mid E = 0, \hat{X}) = 0$ because given $\hat{X}$ and also that no error has occurred, X is no longer random. The second equation is the chain rule (see property (2.6), here we have its conditional form)

$$\begin{aligned} H(X, E \mid \hat{X}) &= H(X \mid \hat{X}) + H(E \mid \hat{X}, X) \\ &= H(X \mid \hat{X}) \end{aligned}$$

where we used the fact that given $\hat{X}$ and X, E is completely determined. Next, we combine these two equations and bound some terms to obtain the desired inequality

$$\begin{aligned} H(X \mid \hat{X}) &= \mathbb{P}(\hat{X} \neq X) H(X \mid E = 1, \hat{X}) + H(E \mid \hat{X}) \\ &\leq \mathbb{P}(\hat{X} \neq X) \log(|\mathcal{X}| - 1) + h_2(\mathbb{P}(\hat{X} \neq X)) \end{aligned}$$

where we used the fact that given $E = 1$, X can have at most $|\mathcal{X}| - 1$ values and the entropy is always bounded from the logarithm of the number of different possible values. ■

If X is distributed uniformly, as in the hypothesis testing, we can further manipulate the previous inequality

Corollary 2.4. *Let $X \to Y \to \hat{X}$ be a Markov chain with X being uniformly distributed on $\mathcal{X}$. We have that*

$$\mathbb{P}(\hat{X} \neq X) \geq 1 - \frac{I(X; Y) + \log 2}{\log |\mathcal{X}|} \tag{2.12}$$

Proof. First we use the definition of mutual information $I(X; \hat{X}) = H(X) - H(X \mid \hat{X})$ and fact that $h_2(p) \leq \log 2$ for every p as the entropy of a binary random variable. Denote by P_e the probability or error $\mathbb{P}(\hat{X} \neq X)$ to save space.

$$\log 2 + P_e \log |\mathcal{X}| \geq h_2(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X \mid \hat{X}) = H(X) - I(X; \hat{X})$$

Next we use the data processing inequality (2.7) which translates to $I(X; \hat{X}) \leq I(X; Y)$. We also use $H(X) = \log |\mathcal{X}|$ because X is uniformly distributed.

$$\log 2 + P_e \log |\mathcal{X}| \geq H(X) - I(X; \hat{X}) \geq \log |\mathcal{X}| - I(X; Y)$$

Therefore, the result follows. ■

Note that above, $\hat{X}$ was an arbitrary estimator. Inequality (2.12) has an intuitive interpretation. A uniform random variable X has entropy $\log|\mathcal{X}|$, which means that one needs logarithmically many bits to describe it. Thus, in order to learn X from Y with small probability of error, Y needs to be such that the mutual information between the two be comparable with $\log|\mathcal{X}|$.

Chapter 3

Estimation Using Many Samples

In this chapter we define the problems examined in this thesis and start developing our results for the case where multiple samples but no queries are allowed. More specifically, we define our noise model and state the problems examined, namely estimation of the adjacency matrix of the unknown graph, exact estimation of edges or triangles with high probability and approximate estimation of edges or triangles with high probability. We derive simple estimators for these problems and analyze their sample complexity. Also, we establish information theoretic lower bounds using the tools of Chapter 2 to show that these estimators are optimal regarding the number of samples they use. This analysis will help in finding an appropriate query strategy to reduce the variance of the estimation, which will be developed in the next chapter.

3.1 The Model

3.1.1 Noise Samples

We begin by formally defining the noise model which we will be using throughout this thesis. According to this model, there exists an undirected graph of interest $G = (V, E)$, where V is a set of n vertices and $E \subseteq V \times V$ is the set of m edges, which we call the *underlying* graph. A *sample* graph G_s of G is a spanning subgraph of G which is generated by the following random process. Each vertex $v \in V$ executes an independent Bernoulli trial with probability of success p. Let $\{X_v\}_{v \in V}$ be those Bernoulli random variables. The edge set E_s of the sample graph contains all edges from G that have at least one endpoint for which the Bernoulli trial failed, that is

$$E_s = \{(u, v) \in E(G) \mid X_u = 0 \vee X_v = 0\}$$

The interpretation of this model is the following. Social networks are undirected graphs which have a vertex for every user and an edge between each pair of friends. Our model describes the process of hiding friends from the public. More specifically, each user decides whether to hide her friends list and if a pair of friends make this decision, their connecting edge becomes invisible to the public. We model the

decision of hiding friends as a random experiment with known probability of success p. We will often call the vertices for which the random trial succeeded *hidden* and the vertices for which it failed *visible*. This graph generating proccedure can be seen as the definition of a probability distribution on all graphs with n vertices. This means that given a graph G we just defined a probability distribution $\mathcal{P}(G)$ according to which each sample graph is generated with a corresponding probability. A generalized version of this model would require a different probability p_v for each user and additionally these probabilities would be unknown, but this would complicate significantly the estimation tasks.

3.1.2 The Oracle

Our algorithms will have access to one or more samples of the form discussed above as well as additional oracle access to G. The oracle responds to a query for a vertex $v \in V$ by returning the neighborhood of that particular vertex in the underlying graph $\Gamma(v) = \{u \in V \mid (u, v) \in E\}$. Clearly, visible vertices already reveal their true neighborhoods in the sample graphs and thus it would be feckless to query for them. However, it is stressed that our algorithms do not have knowledge of the Bernoulli trials vector $\mathbf{X} = (X_{v_1}, \ldots, X_{v_n})$ and thus do not know which are the visible vertices.

3.2 General Objectives

N sample graphs $G_1, \ldots, G_N$ are revealed and the algorithms have additional query access to the underlying graph G. The problems examined are that of estimating parameters of the underlying graph G. In this thesis we will mainly focus on the following parameters.

1. **Adjacency matrix or graph estimation**. Given a confidence parameter $\delta \in (0, 1]$ the goal is to find a graph $\hat{G}$ with n vertices such that

$$\mathop{\mathbb{P}}_{G_1,\ldots,G_N \sim \mathcal{P}_G}(\hat{G} = G) \geq 1 - \delta \tag{3.1}$$

2. **Exact edge estimation**. Given $\delta \in (0, 1]$ the goal is to find an estimator $\hat{m}$ for the edges of the underlying graph $m = |E(G)|$ such that

$$\mathop{\mathbb{P}}_{G_1,\ldots,G_N \sim \mathcal{P}_G}(\hat{m} = m) \geq 1 - \delta \tag{3.2}$$

3. **Exact triangle estimation**. The same for estimating the number of triangles. This can be generalized to any arbitrary subgraph occurrences estimation.

4. **Approximate edge estimation**. For every $\varepsilon, \delta \in (0, 1]$ we seek to find an (ε, δ)-estimator, that is an estimator $\hat{m}$ for the edges of the underlying graph $m = |E(G)|$ so that the relative error is small with high probability

$$\mathop{\mathbb{P}}_{G_1,\ldots,G_N \sim \mathcal{P}_G}(|\hat{m} - m| \leq \varepsilon m) \geq 1 - \delta \tag{3.3}$$

5. **Approximate triangle estimation**. The same for triangles.

For all these estimation tasks we will be interested in the sample complexity which is the number of samples required to obtain the stated guarantees and the query complexity which is the number of queries required. Essentially we would like to study the trade off between sample complexity and query complexity, as it is intuitively apparent that samples reveal a lot of information which may be redundant while queries have the potential to reveal less but more relevant information. Thus, with a few sample graphs we should be able to learn the graph well enough in order to decide which vertices are critical for the property under examination.

3.3 Learning the Underlying Graph

We start our study by examining some simple estimators for the problems listed above, which use only samples of the underlying graph and no queries. This study gives some insight about the sample complexity of the estimation problems mentioned above, as well as ideas on how to improve them by allowing vertex queries, which will be done in later chapters.

3.3.1 Maximum Likelihood Estimator

First, we examine an estimator of the whole underlying graph. Such an estimator needs to determine the edge set of G, with high probability. Every sample is a subgraph of the underlying graph, meaning that every edge observed is a real edge of the underlying graph while every pair of vertices with no observed edge between them could potentially have an edge in G. Consider the simple estimator of taking the union of all sample graphs $G_1, \ldots, G_N$. This seems like a very natural choice which is supported by the following proposition.

Proposition 3.1. *The union estimator is a maximum likelihood estimator for the underlying graph G.*

$$(\hat{G})_{ML} = \bigcup_{i=1}^{N} G_i$$

Proof. For the proof we need to argue that including more edges to $\hat{G}$ than those observed in the samples can only reduce the likelihood. Because of the fact that n is fixed, a graph is solely described by its the pairs of vertices connected by edges. For a sequence of graphs $g_1, \ldots, g_N$ let $\mathbf{e_i} = (e_i^1, \ldots, e_i^N)$ be a binary vector that describes the edges of those graphs as following: each element e_i^j is one if and only if the i-th pair of vertices is connected by an edge in the j-th graph, where $i \in [n(n-1)/2]$ and $j \in [N]$. With capital letter we will denote the same vectors regarding the observed sample graphs. Next we define the likelihood function $\mathcal{L}(g_1, \ldots, g_N \,;\, G)$ and show that it is maximized when $G = \cup_{i=1}^{N} g_i$.

$$\begin{aligned}\mathcal{L}(g_1, \ldots, g_N \,;\, G) &= \mathbb{P}(G_1 = g_1, \ldots, G_N = g_N \mid G) \\ &= \mathbb{P}(\mathbf{E}_1 = \mathbf{e}_1, \ldots, \mathbf{E}_{n(n-1)/2} = \mathbf{e}_{n(n-1)/2} \mid G)\end{aligned}$$

$$= \prod_{i=1}^{n(n-1)/2} \mathbb{P}(\mathbf{E}_i = \mathbf{e}_i \mid \mathbf{E}_{i-1} = \mathbf{e}_{i-1}, \ldots, \mathbf{E}_1 = \mathbf{e}_1, G)$$

We need to choose G such that it does not miss any edges from those of $\cup_{i=1}^{N} g_i$, because otherwise the likelihood becomes zero. In addition, suppose we include in G an edge that has not been observed in any sample, say we include the i-th pair. Then the corresponding factor from the product above $\mathbb{P}(\mathbf{E}_i = \mathbf{e}_i \mid \mathbf{E}_{i-1} = \mathbf{e}_{i-1}, \ldots, \mathbf{E}_1 = \mathbf{e}_1, G)$ will be less than or equal to one, where if we had not included this edge, the factor would be equal to one. This shows that the likelihood may drop if we choose more edges than those of the union and concludes the proof. ■

After we showed that the union is a maximum likelihood estimator, we would like to find how many samples it takes for that estimator to obtain low probability of error. The way this estimator works is that by taking many samples, the chances of each vertex to become visible in some of these samples increase and if the number of samples is great enough, then all vertices will show their true neighborhoods. A simple analysis shows that $\Theta(\log n)$ samples suffice for that purpose.

Proposition 3.2. *Let $\delta \in (0, 1]$. If the number of samples is $N > \log(n/\delta)$, it holds for the union estimator $\hat{G} = \cup_{i=1}^{N} G_i$ that*

$$\mathbb{P}_{G_1,\ldots,G_N \sim \mathcal{P}_G}(\hat{G} = G) \geq 1 - \delta$$

Proof. We bound the probability of wrong estimation. Denote by $\{X_u\}_{u \in V}$ the Bernoulli random variables associated with the vertices of the graph.

$$\begin{aligned}
\mathbb{P}_{G_1,\ldots,G_N \sim \mathcal{P}_G}(\hat{G} \neq G) &\leq \mathbb{P}_{G_1,\ldots,G_N \sim \mathcal{P}_G}[\exists u \in V \ : \ X_u^1 = 1, \ldots, X_u^N = 1] \\
&\leq \sum_{u \in V} \mathbb{P}_{G_1,\ldots,G_N \sim \mathcal{P}_G}(X_u^1 = 1, \ldots, X_u^N = 1) \\
&= np^N \\
&< \delta
\end{aligned}$$

where we used union bound for the second inequality. The last inequality comes from the fact that $N > \log(n/\delta)$. ■

From the proof above it becomes more clear why a union of many samples is more representative than one sample, more specifically *how* it is improved as the number of samples increases. This is an observation that will be used later.

Remark 3.1. *The union of many, say N, samples can be considered as a single sample generated by the same noise model, but with smaller probability for the vertices to be hidden, specifically with $p' = p^N$.*

3.3.2 Lower Bound

Now that we provided a first estimator for the adjacency matrix of the underlying graph and we upper bounded the sample complexity of this problem, the next natural step towards getting insight about the problem is to examine the optimality

of that estimator, before seeking for more advanced ones. Also we hope that through that study we will be able to answer questions regarding the sample complexity of the estimating problems stated at the start of the chapter, something that will be discussed later in this chapter.

Definition 3.1. *The sample complexity of estimating the adjacency matrix is a function $f(n, \delta)$ of the size n of the underlying graph G and the confidence parameter δ, such that every algorithm $\hat{G}$ for this task requires at least $f(n)$ samples in order to satisfy*

$$\forall G \, \mathbb{P}(\hat{G} = G) \geq 1 - \delta$$

This definition is generalised for every other estimation task in this thesis. For simplicity we will only care about the dependence on the size of the graph, equivalently we may replace δ with a constant, say $1/3$ in the definition above. To determine the sample complexity of learning the underlying graph, we first adopt a more intuitive and informal approach that hopefully reveals the idea behind the lower bound construction and then we give a formal proof using Fano's inequality and the tools presented in Chapter 2.

Consider the class of graphs with only two vertices. Let G_1 be the graph that has the edge and G_0 the graph with the missing edge. Define the following hypothesis testing: A graph is chosen between G_0 and G_1 uniformly at random and is used to generate N sample graphs $\mathbf{X} = (X_1, \ldots, X_N)$. An estimator $\hat{G}$ that has access to these samples solves the problem of hypothesis testing if it correctly recovers the decision made, that is, which graph was used to generate the samples. The estimator may be randomized. This means that for fixed input samples the estimator produces a probability mass over the two graphs. The samples are independent, so without loss of generality we can assume that the estimation does not depend on the order of the observations. In other words, if we change the order of the input samples, the probability mass that the estimator outputs should be the same. Additionally, if we want to consider only useful estimators we can make the assumption that if one of the samples contains an edge the estimator gives G_1 as the answer with probability one. The probability of correct estimation is

$$\frac{1}{2} \mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{P}_{G_1}^N} (\hat{G}(\mathbf{X}) = G_1) + \frac{1}{2} \mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{P}_{G_0}^N} (\hat{G}(\mathbf{X}) = G_0)$$

We examine the two terms separately, (a will be a parameter of the estimator to model its randomization). For the first term we have

$$\begin{aligned} \mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{P}_{G_1}^N} \left(\hat{G}(\mathbf{X}) = G_1 \right) &= \mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{P}_{G_1}^N} \left(\hat{G}(\mathbf{X}) = G_1 \mid \bigcup_i X_i = G_1 \right) \mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{P}_{G_1}^N} \left(\bigcup_i X_i = G_1 \right) \\ &\quad + \mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{P}_{G_1}^N} \left(\hat{G}(\mathbf{X}) = G_1 \mid \bigcup_i X_i = G_0 \right) \mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{P}_{G_1}^N} \left(\bigcup_i X_i = G_0 \right) \\ &= 1 \cdot (1 - p^{2N}) + a p^{2N} \end{aligned} \tag{3.4}$$

For the second term we have

$$\mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{P}_{G_0}^N} (\hat{G}(\mathbf{X}) = G_0) = \mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{P}_{G_0}^N} \left(\hat{G}(\mathbf{X}) = G_0 \mid \bigcup_i X_i = G_1 \right) \mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{P}_{G_0}^N} \left(\bigcup_i X_i = G_1 \right)$$

$$+ \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}^N_{G_0}}\left(\hat{G}(\mathbf{X}) = G_0 \mid \bigcup_i X_i = G_0\right) \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}^N_{G_0}}\left(\bigcup_i X_i = G_0\right)$$
$$= 0 + (1-a) \tag{3.5}$$

Figure 3.1. Two graphs of the family used for the lower bound when $n = 8$. The left is $G_{10100111}$ and the right is $G_{00111011}$.

Now consider the family of graphs (see Figure 3.1) $\{G_v\}_{v\in\mathcal{V}}$ with n disjoint pairs of vertices that can have an edge between each one of these pairs (2^n graphs in total). This family is the cartestian product of families like the one defined previously. Let $\mathcal{V} = \{0,1\}^n$ be the set of indexes for the graphs of this family (a zero in the index of the graph means that the edge associated with that particular position is missing, where a one denotes an existing edge). The hypothesis testing is the following set up: An index from $\mathcal{V}$ is picked uniformly at random according to a random variable V. Conditioned on the choice $V = v$, N samples are drawn from the distribution of G_v. The goal is to determine the value v of the index V. Let $\mathbf{\Psi} = (\Psi_1, \ldots, \Psi_n)$ be an arbitrary estimator for the hypothesis testing for this family (it is a n-dimensional binary vector which has an element for each possible edge). The edges of these graphs are disjoint, so without loss of generality we can assume that the estimator operates independently on each pair of vertices, like the estimator described above. In other words, we can assume that each Ψ_i is an estimator like the one described previously and operates independently from the others. We have the following for the probability of error (to avoid complicated notation we will drop subscripts and were no confusion can occur)

$$\begin{aligned}
\mathbb{P}(\mathbf{\Psi}(\mathbf{X}) \neq G_V) &= 1 - \mathbb{P}(\mathbf{\Psi}(\mathbf{X}) = G_V) \\
&= 1 - \sum_{v\in\{0,1\}^n} \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}^N_{G_v}} (\mathbf{\Psi}(\mathbf{X}) = G_V \mid V = v)\, \mathbb{P}(V = v) \\
&= 1 - \frac{1}{2^n} \sum_{(e_1,\ldots,e_n)\in\{0,1\}^n} \mathbb{P}(\mathbf{\Psi}(\mathbf{X}) = G_{(e_1,\ldots,e_n)}) \\
&= 1 - \frac{1}{2^n} \sum_{(e_1,\ldots,e_n)\in\{0,1\}^n} \mathbb{P}(\Psi_1(\mathbf{X}) = e_1, \ldots, \Psi_n(\mathbf{X}) = e_n) \\
&= 1 - \frac{1}{2^n} \sum_{e_1\in\{0,1\}} \cdots \sum_{e_n\in\{0,1\}} \mathbb{P}(\Psi_1(\mathbf{X}) = e_1) \cdots \mathbb{P}(\Psi_n(\mathbf{X}) = e_n) \\
&\qquad\qquad \text{(independence)} \\
&= 1 - \frac{1}{2^n} \sum_{e_1\in\{0,1\}} \mathbb{P}(\Psi_1(\mathbf{X}) = e_1) \cdots \sum_{e_n\in\{0,1\}} \mathbb{P}(\Psi_n(\mathbf{X}) = e_n)
\end{aligned}$$

$$= 1 - \frac{1}{2^n}((1-p^{2N}) + ap^{2N} + 1 - a)^n \qquad \text{(from 3.4, 3.5)}$$
$$= 1 - \left(\frac{2 - a - (1-a)p^{2N}}{2}\right)^n$$

It is now a matter of analysis to show that this error is great if fewer than logarithmic samples are used. For the sake of completeness we present a proof.

Lemma 3.1. *For every* $a \in [0,1]$*, if* $N \leq \frac{1}{2}\log\frac{n}{2\ln(3/2)}$*, then* $2^{-n}(2 - a - (1-a)p^{2N})^n \leq 2/3$.

Proof. Define the function under examination

$$f(N, n, a; p) = \left(\frac{2 - a - (1-a)p^{2N}}{2}\right)^n$$

We begin by examining the case $a = 0$, because this is the case were the estimator makes most sense intuitively ($a = 0$ means that the estimator does not output more edges than those observed in the samples, just like the MLE estimator of the previous section). Substituting $a = 0$ gives, assuming $N \leq \frac{1}{2}\log\frac{n}{2\ln(3/2)}$, that

$$f(N, n, a = 0; p) = \left(\frac{2 - p^{2N}}{2}\right)^n = \left(1 - \frac{1}{2}p^{2N}\right)^n \leq \exp\left(-\frac{1}{2}np^{2N}\right) \leq \frac{2}{3}$$

Next we show that if $a > 0$ this quantity is even smaller than the previous case. To see this differentiate with respect to a and observe that the derivative is negative

$$\frac{\partial}{\partial a}f(N, n, a; p) = \frac{n}{2^n}(p^{2N} - 1)(2 - a + (a-1)p^{2N})^{n-1} < 0$$

■

We have thus shown that if $N = o(\log n)$, it holds

$$\mathbb{P}(Error) \geq 1/3$$
$$\mathbb{P}(\mathbf{\Psi}(\mathbf{X}) \neq G_V) \geq 1/3$$
$$\frac{1}{2^n}\sum_v \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}^N_{G_v}}(\mathbf{\Psi}(\mathbf{X}) \neq G_v) \geq 1/3$$
$$\exists v \in \mathcal{V} \quad \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}^N_{G_v}}(\mathbf{\Psi}(\mathbf{X}) \neq G_v) \geq 1/3$$

Noting that the estimator $\mathbf{\Psi}$ was arbitrary, we have shown that for every estimator there exists a graph such that the estimator fails to recognize it with constant probability.

3.3.3 Fano's Inequality

Here we prove the same lower bound formally using Fano's inequality. This is a method that costructs lower bounds that are expressed using information theoretic measures, which we will need to calculate. After that, our efforts will be reduced to bounding an analytic expression.

Theorem 3.1. *There exists an estimator $\hat{G}$ such that for every parameter $\delta \in (0,1]$ and graph G, given $N = \Theta(\log(n/\delta))$ i.i.d. samples $X_1, \ldots, X_N \sim \mathcal{P}_G$ satisfies $\mathbb{P}_{\mathbf{X}\sim\mathcal{P}_G^N}(\hat{G} \neq G) < \delta$. In addition, if $N = o(\log n)$ then for every estimator $\hat{G}$ there exists a graph G such that $\mathbb{P}_{\mathbf{X}\sim\mathcal{P}_G^N}(\hat{G} \neq G) \geq 1/3$.*

Proof. The family of graphs $\mathcal{G} = \{G_v\}_{v\in\mathcal{V}}$ we consider now is smaller than the one used in the previous section. For each index $v = \{1, \ldots, n\}$ (not to be confused with a vertex) we define the graph G_v of the family as follows. The vertex set is $\{u_1, \ldots, u_n\} \cup \{w_1, \ldots, w_n\}$ and the edge set is $\{(u_1, w_1), \ldots, (u_n, w_n)\} \setminus \{(u_v, w_v)\}$. Again, let $\mathcal{V}$ be the set of indices of the graphs and V the random variable which denotes which one of the n graph distributions was initially chosen. Thus the hypothesis testing problem we define consists of the following: First, an index V from $\mathcal{V}$ is chosen uniformly at random and then N sample graphs $G_1, \ldots, G_N$ or, more compactly, $\mathbf{G}$ are drawn from the distribution defined by the underlying graph G_V. The estimator Ψ has to output the value of V correctly, having access to only the sample graphs. Following the notation of Chapter 2, we denote by $I(V; G_1, \ldots, G_N)$ or $I(V; \mathbf{G})$ the mutual information between the samples and the random variable V. The inequality we are using is (2.12) which we restate here.

$$\mathbb{P}(\Psi(\mathbf{G}) \neq G_V) \geq 1 - \frac{I(V;\mathbf{G}) + \log 2}{\log|\mathcal{V}|} \tag{3.6}$$

We denote by $\mathcal{P}_{V,\mathbf{G}}$ the joint distribution of V and the samples, by $\mathcal{P}_V$ the distribution of V which is uniform and by $\mathcal{P}_{\mathbf{G}} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathcal{P}_{G_v}^N$ the joint distribution of the samples which is a mean of the distributions defined by the graphs of the family. We calculate the mutual information and provide some explanations about these calculations below:

$$\begin{aligned}
I(V;\mathbf{G}) &= D_{\mathrm{KL}}\left(\mathcal{P}_{V,\mathbf{G}} \middle|\middle| \mathcal{P}_V\mathcal{P}_{\mathbf{G}}\right) \\
&= \sum_{v\in\mathcal{V},\mathbf{g}\in\mathcal{G}^N} \mathbb{P}_{(V,\mathbf{G})\sim\mathcal{P}_{V,\mathbf{G}}}(\mathbf{G}=\mathbf{g}, V=v) \log\left(\frac{\mathbb{P}_{(V,\mathbf{G})\sim\mathcal{P}_{V,\mathbf{G}}}(\mathbf{G}=\mathbf{g}, V=v)}{\mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{\mathbf{G}}}(\mathbf{G}=\mathbf{g})\,\mathbb{P}_{V\sim\mathcal{P}_V}(V=v)}\right) \\
&= \sum_{v\in\mathcal{V}} \mathbb{P}_{V\sim\mathcal{P}_V}(V=v) \sum_{\mathbf{g}\in\mathcal{G}^N} \mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{G_v}^N}(\mathbf{G}=\mathbf{g}) \log\left(\frac{\mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{G_v}^N}(\mathbf{G}=\mathbf{g})}{\mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{\mathbf{G}}}(\mathbf{G}=\mathbf{g})}\right) \\
&= \frac{1}{n}\sum_{v\in\mathcal{V}}\sum_{\mathbf{g}\in\mathcal{G}^N} \mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{G_v}^N}(\mathbf{G}=\mathbf{g}) \log\left(\frac{\mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{G_v}^N}(\mathbf{G}=\mathbf{g})}{\frac{1}{n}\sum_{v'\in\mathcal{V}}\mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{G_{v'}}^N}(\mathbf{G}=\mathbf{g})}\right) \\
&= \frac{1}{n}\sum_{v\in\mathcal{V}}\sum_{g\in\mathcal{G}} \mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{G_v}^N}\left(\bigcup_i G_i = g\right) \log\left(\frac{\mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{G_v}^N}(\cup_i G_i = g)}{\frac{1}{n}\sum_{v'\in\mathcal{V}}\mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{G_{v'}}^N}(\cup_i G_i = g)}\right) \\
&= \frac{1}{n} n \sum_{g\in\mathcal{G}} \mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{G_v}^N}\left(\bigcup_i G_i = g\right) \log\left(\frac{\mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{G_v}^N}(\cup_i G_i = g)}{\frac{1}{n}\sum_{v'\in\mathcal{V}}\mathbb{P}_{\mathbf{G}\sim\mathcal{P}_{G_{v'}}^N}(\cup_i G_i = g)}\right) \\
&= \sum_{k=0}^{n-1}\binom{n-1}{k} p^{2Nk}(1-p^{2N})^{n-1-k} \log\left(\frac{n}{k+1}\right)
\end{aligned}$$

Where, for the fifth equality we did not assumed that $\mathbb{P}(\mathbf{G} = \mathbf{g}) = \mathbb{P}(\cup_i G^i = g_i)$ as it may seem (this is not generaly true). Instead, one needs to think that for all sample graph vectors $\mathbf{g} = (g^1, \dots, g^N)$ with the same union, the argument inside the logarithm is the same and thus we can group these terms together (the argument inside the logarithm is n over the number of graphs in the family that can generate the samples with non zero probability). For the next equality the terms for every $v \in \mathcal{V}$ are equal due to symmetry reasons and thus the first sum degenerates to n equal terms. For the last equality, we group all possible graphs according to the number of hidden edges. Thus Fano's inequality (3.6) eventually has been turned into

We now need to show that if $N = o(\log n)$, the quantity $I(V; \mathbf{G})/\log n$ goes to zero. Observe that the sum can be treated as an expected value involving a binomial random variable $X \sim Bin(n-1, p^{2N})$ and use Taylor expansion to approximate it.

$$\sum_{k=0}^{n-1} \binom{n-1}{k} p^{2Nk}(1-p^{2N})^{n-1-k} \log\left(\frac{n}{k+1}\right) = \mathbb{E}\left[\log\left(\frac{n}{X+1}\right)\right]$$

Taylor approximation for calculating expectations. Let X be a random variable and μ_X its expectation. We can approximate the expected value of a function of X as following:

$$\begin{aligned}\mathbb{E}[f(X)] &= \mathbb{E}[f(\mu_X + (X - \mu_X))] \\ &\approx \mathbb{E}\left[f(\mu_X) + f'(\mu_X)(X-\mu_X) + \frac{1}{2}f''(\mu_X)(X-\mu_X)^2\right] \\ &= f\left(\mathbb{E}[X]\right) + \frac{1}{2}f''\left(\mathbb{E}[X]\right)\mathrm{Var}[X] \end{aligned} \tag{3.7}$$

Applying this to our case, $\mathbb{E}[X] = (n-1)p^{2N}$, $\mathrm{Var}[X] = (n-1)p^{2N}(1-p^{2N})$, and

$$\begin{aligned} f(x) &= \log\left(\frac{n}{x+1}\right) \\ f'(x) &= -\frac{1}{x+1} \\ f''(x) &= \frac{1}{(1+x)^2} \end{aligned}$$

Therefore we have that

$$\frac{1}{\log n}\mathbb{E}\left[\log\left(\frac{n}{X+1}\right)\right] \approx \frac{1}{\log n}\log\left(\frac{n}{1+(n-1)p^{2N}}\right) + \frac{1}{2\log n}\frac{(n-1)p^{2N}(1-p^{2N})}{[1+(n-1)p^{2N}]^2}$$

Now it easy to see that if $N = o(\log n)$ both terms become $o(1)$. According to (3.6), this means that the error $\mathbb{P}(\Psi(\mathbf{G}) \neq G_V)$ goes to one as the size n increases. Therefore, we have that for every estimator there exists a graph such that the error, when samples $o(\log n)$ from this graph distribution are used, is high.

$$\begin{aligned} \mathbb{P}(\mathbf{\Psi}(\mathbf{X}) \neq G_V) &\geq 1/3 \\ \frac{1}{n}\sum_v \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{P}^N_{G_v}}(\mathbf{\Psi}(\mathbf{X}) \neq G_v) &\geq 1/3 \end{aligned}$$

$$\exists v \in \mathcal{V} \qquad \mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{P}_{G_v}^N} (\mathbf{\Psi}(\mathbf{X}) \neq G_v) \geq 1/3$$

■

We conclude with some remarks. The idea behind the lower bound essentially is that the union of the samples provides all the necessary information about the underlying graph. This was made intuitively understandable by the arguments provided in the begining of the section and finally formalized in the information theoretic proof, where we showed that the mutual information between the underlying graph and the sample graphs is the same as the mutual information between the underlying graph and the union of the samples. The other observation is that the family of graphs used to prove the bound was not really a graph with many connections between its vertices but rather a collection of gadgets that served as bits of information.

Remark 3.2. *Learning a word of n bits using noisy samples, were each bit that equals to one may be zeroed in the samples with some fixed probability, needs $\Omega(\log n)$ samples.*

3.4 Exact Edge and Triangle Estimation

Having studied the graph learning problem we move to the estimation problems of main interest in this thesis, which are counting the occurrences of fixed subgraphs in the underlying graph. It is stressed that in this section we seek for estimators satisfying the guarantee (3.2). We again intend to start from simple estimators for each problem, examine their sample complexities and compare them with the sample complexity of the estimation problems. It may seem plausible at first that, depending on the property of interest, its estimation will have different sample complexity. For example, learning the number of edges seems less demanding than the number of triangles and the latter seems less demanding than estimating the number of circles of length 5 or the diameter of the graph. Thus, through our study of sample complexity, we would like to approach the following problem.

Question 3.1. *Is the sample complexity of estimating global properties higher than the sample complexity of estimating local properties?*

Or perhaps the more ambitious question:

Question 3.2. *Can the sample complexity of learning properties of graphs characterize them in any way?*

Our analysis gives a negative answer to those questions as we again derive lower bounds that are logarithmic in the number of vertices of the graph, which are the same as the bound for learning the whole graph. This means that, at least for the properties examined here, their exact estimation is not easier than learning the graph.

3.4.1 Upper Bound

One estimator for counting subgraphs exactly would be to use the union estimator and return the number of edges observed in the union of the samples. With $\Theta(\log(n/\delta))$ samples, the structure of the graph can be learned with probability at least $1-\delta$. which provides a naive estimator for every graph property. It remains to show that the lower bound matches this upper bound.

3.4.2 Lower Bound

We derive a statistical lower bound, again using the machinery presented in Chapter 2. We search for the family of graphs for the construction of the bound by searching for the most difficult input for *every* estimator. In other words, we want to find graphs that have different value of the property of interest but are similar and thus hard to be distinguished.

Theorem 3.2. *There exists an estimator $\hat{m}$ such that for every $\delta \in (0,1]$ and graph G, given $N = \Theta(\log(n/\delta))$ i.i.d. samples $X_1, \ldots, X_N \sim \mathcal{P}_G$ satisfies $\mathbb{P}_{\mathbf{X}\sim\mathcal{P}_G^N}(\hat{m} \neq |E(G)|) < \delta$. In addition, if $N = o(\log n)$ then for every estimator $\hat{m}$ there exists a graph G such that $\mathbb{P}_{\mathbf{X}\sim\mathcal{P}_G^N}(\hat{m} \neq |E(G)|) \geq 1/3$.*

Proof. For this proof, we use a binary hypothesis testing. Let H_0 be the null hypothesis and H_1 the alternative hypothesis. We describe the graphs associated with each hypothesis. For H_0 the underlying graph is a matching. More specifically, its vertex set is $\{u_1, \ldots, u_n\} \cup \{v_1, \ldots, v_n\}$ and its edge set is $\{(u_1, v_1), \ldots, (u_n, v_n)\}$.

H_1 is a composite hypothesis, that is, in this case a graph is selected uniformly at random from the family of graphs $\{G_i\}_{i=1}^n$ where the i-th graph is a matching, like the graph of the null hypothesis, but with the i-th edge missing (this is the same family used in the previous lower bound).

Each hypothesis is selected with probability $1/2$ according to a random variable V. If there exists an estimator $\hat{m}$ that guarantees $\mathbb{P}(\hat{m} \neq m) < \delta$ with $N = o(\log n)$ samples then we could use it to test the hypotheses presented above with error probability less than δ. However we will show that solving the hypothesis testing with arbitrarily small probability of error, say $1/3$, requires $\Omega(\log n)$ samples and thus the edge estimator mentioned before cannot exist.

Let $\mathcal{P}_1$ be the joint distribution of the samples defined by the graph of the null hypothesis and $\mathcal{P}_2$ be the joint distribution of the samples if the alternative hypothesis is selected. Also, let Ψ be an arbitrary estimator for this hypothesis testing. From Le Cam (2.10) and Pinsker (2.5) we have the bound:

$$\begin{aligned}
\mathbb{P}(\Psi \neq V) &= \frac{1}{2}\mathop{\mathbb{P}}_{H_0}(\Psi \neq 0) + \frac{1}{2}\mathop{\mathbb{P}}_{H_1}(\Psi \neq 1) \\
&= \frac{1}{2}\left(1 - \|\mathcal{P}_2 - \mathcal{P}_1\|_{TV}\right) \\
&\geq \frac{1}{2}\left(1 - \sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathcal{P}_2 || \mathcal{P}_1\right)}\right)
\end{aligned} \tag{3.8}$$

It remains to calculate the KL-divergence between the joint distributions. For the alternative hypothesis, observe that the corresponding distribution $\mathcal{P}_2$ is a mean over

the distributions defined by the family $\{G_i\}_{i=1}^n$ because a graph of this family is chosen uniformly at random to generate the N samples. Thus $\mathcal{P}_2 = \frac{1}{n}\sum_{i=1}^n \mathcal{P}_{G_i}^N$. Let $g_1, \ldots, g_N$ be N graphs that belong in the support of the two graph distributions, that is, they are matchings with a number of edges that could be ranging from 0 to $n-1$ (because a graph with n edges can be generated only by $\mathcal{P}_1$). If k of the n possible edges are missing from all $g_1, \ldots, g_N$, the probability each distribution gives to this sequence is

$$P_1(g_1, \ldots, g_N) = p^{2Nk}(1-p^{2N})^{n-k}$$
$$P_2(g_1, \ldots, g_N) = \frac{k}{n} p^{2N(k-1)}(1-p^{2N})^{n-k}$$

where for the second distribution the factor $1/n$ exists because the underlying graph is selected uniformly at random from a family with n graphs and the factor k is there because k graphs of this family can generate the sequence (each with equal probability). The KL-divergence is then

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathcal{P}_2 || \mathcal{P}_1\right) &= \sum_{g_1,\ldots,g_N} P_2(g_1, \ldots, g_N) \log\left(\frac{P_2(g_1, \ldots, g_N)}{P_1(g_1, \ldots, g_N)}\right) \\ &= \sum_{k=1}^{n} \binom{n}{k} \frac{k}{n} p^{2N(k-1)}(1-p^{2N})^{n-k} \log\left(\frac{k}{np^{2N}}\right) \end{aligned}$$

where for the last equality we grouped the vectors $\mathbf{g} = (g_1, \ldots, g_N)$ according to the number of edges k missing from the union $\cup_{i=1}^N g_i$. Thus the lower bound (3.8) becomes

$$\mathbb{P}(\Psi \neq V) \geq \frac{1}{2}\left[1 - \sqrt{\frac{1}{2}\sum_{k=1}^{n} \binom{n}{k} \frac{k}{n} p^{2N(k-1)}(1-p^{2N})^{n-k} \log\left(\frac{k}{np^{2N}}\right)}\right]$$

Again, the construction of the lower bound reduced to bounding an analytic expression. To see that the series go to zero with n if N is less than logarithmic, observe that the sum can be written as the expected value of a Binomial random variable $X \sim Bin(n, p^{2N})$, more specifically, we can write

$$\sum_{k=1}^{n} \binom{n}{k} \frac{k}{n} p^{2N(k-1)}(1-p^{2N})^{n-k} \log\left(\frac{k}{np^{2N}}\right) = \frac{1}{np^{2N}}\,\mathbb{E}\left[X \log\left(\frac{X}{np^{2N}}\right)\right]$$

For the random variable X we have that $\mathbb{E}[X] = np^{2N}$ and $\mathrm{Var}[X] = np^{2N}(1-p^{2N})$. We approximate the above expectation by keeping the first two terms of the Taylor expansion, that is, we use (3.7) where

$$f(x) = x \log\left(\frac{x}{np^{2N}}\right)$$
$$f''(x) = \frac{1}{x}$$

From this process we get the following

$$\frac{1}{np^{2N}}\,\mathbb{E}\left[X \log\left(\frac{X}{np^{2N}}\right)\right] \approx \frac{np^{2N}}{np^{2N}} \log\left(\frac{np^{2N}}{np^{2N}}\right) + \frac{1}{2}\frac{np^{2N}(1-p^{2N})}{(np^{2N})^2}$$

$$= 0 + \frac{1 - p^{2N}}{2np^{2N}} \xrightarrow[N=o(\log n)]{n\to\infty} 0$$

Therefore, we showed that if $N = o(\log n)$ samples are used, every estimator for the hypothesis testing has probability of error that goes to $1/2$. This means that for big enough n, say $n \geq n_0$, it holds that $\mathbb{P}(\Psi \neq V) \geq 1/3$ and because $\mathbb{P}(\Psi \neq V) = \frac{1}{2}\mathbb{P}_{H_0}(\Psi \neq 0) + \frac{1}{2}\mathbb{P}_{H_1}(\Psi \neq 1)$, we get that for some $i = \{0, 1\}$ it is true that $\mathbb{P}_{H_i}(\Psi \neq i) \geq 1/3$ which concludes the proof. ■

An observation, relevant to the previous bound about learning the underlying graph is that then we examined how many samples it takes to learn the labeled graph. Here we derived that the same holds true about learning the unlabeled graph.

Corollary 3.1. *The sample complexity of learning the unlabeled underlying graph is* $\Theta(\log n)$.

3.4.3 Triangle Estimation

The previous lower bound can be directly generalized for triangle estimation or even arbitrary subgraph occurrences estimation. The reason is that the underlying graphs used in the proof consisted of disjoint gadgets serving as bits of information. Instead of using edges as those gadgets, we can use triangles or other small fixed graphs and obtain the same bounds for the corresponding estimation problems. Thus, we do not repeat the full proof here but we only point out the changes needed to generalize it.

The hypothesis testing for the triangles is the following. The graph for the null hypothesis H_0 has vertex set $\{u_1, \ldots, u_n\} \cup \{v_1, \ldots, v_n\} \cup \{w_1, \ldots, w_n\}$ and edge set $\{(u_i, v_i)\}_{i=1}^n \cup \{(v_i, w_i)\}_{i=1}^n \cup \{(w_i, u_i)\}_{i=1}^n$.

The alternative hypothesis H_1 is again a composite hypothesis, where a graph is selected uniformly at random from the family of n graphs $\{G_i\}_{i=1}^n$ where the i-th graph is defined like the graph of the null hypothesis with the only difference that it misses the i-th triangle (it misses all three edges of that particular triangle).

If we had a triangle estimator $\hat{T}$ such that $\mathbb{P}(\hat{T} \neq T) < \delta$, we could use it to test which hypothesis holds with probability of error at most δ but we will show that the former task requires $\Omega(\log n)$ samples to achieve arbitrarily small probability of error. Analogously to the case of edges, let $\mathcal{P}_1, \mathcal{P}_2$ be the *joint distributions* of the samples conditioned on each hypothesis. The only change in the calculation of $D_{\mathrm{KL}}(\mathcal{P}_2 || \mathcal{P}_1)$ now is the fact that a triangle becomes entirely hidden with probability p^3 instead of p^2 and therefore, for a graph sequence $g_1, \ldots, g_N$ such that $\cup_{i=1}^N g^i$ has k missing triangles (we will call a triangle missing if all of its edges are missing) the two distributions produce it with probabilities

$$P_1(g^1, \ldots, g^N) = p^{3Nk}(1 - p^{3N})^{n-k}$$
$$P_2(g^1, \ldots, g^N) = \frac{k}{n} p^{3N(k-1)}(1 - p^{3N})^{n-k}$$

and the proof continues with the only difference $p^2 \leftrightarrow p^3$.

Remark 3.3. *Estimating the number of squares or, in general, appearances of a fixed subgraph inside the underlying graph has the same sample complexity for the same reason.*

Remark 3.4. *We actually proved that learning the number of triangles in the complementary graph needs* $\Omega(\log n)$ *samples.*

3.5 Approximate Estimation of Edges

In this section, we are interested in approximating the number of edges m of the underlying graph. More specifically, we would like to have an estimator $\hat{m}$ such that $(1-\varepsilon)m \leq \hat{m} \leq (1+\varepsilon)m$ with probability at least $1-\delta$. While, the main objective of this thesis is to examine triangles or other induced subgraphs, we start from edge estimation mainly for two reasons. The first one is that the number of edges is more closely related to the degrees of the graph, which are somewhat preserved in the samples (the degree of each vertex is either the same with its degree in the underlying graph or follows a binomial distribution) and thus edge estimation is more natural to start with. The second reason is that we will develop a unified approach for all these estimation problems based on the simpler case of the edges. Again, we note that thought this chapter we will be considering estimators having access only to samples of the underlying graph.

3.5.1 Mean and Variance

Let $G = (V, E)$ be the underlying graph, denote by m the number of edges of the underlying graph $|E(G)|$ and by G_s the sample graph. Denote by $\{X_v\}_{v \in V}$ the set of independent Bernoulli random variables with probability of success p, associated with each vertex. $X_v = 1$ means that vertex v decided to hide his neighborhood in the sample graph. We define a random variable $Y_{(u,v)} = X_u X_v$ associated with each edge $(u, v) \in E$. Similarly $Y_{(u,v)} = 1$ if the edge (u, v) is hidden in the sample graph (which happens with probability p^2) and $Y_{(u,v)} = 0$ otherwise. These random variables are not independent if they correspond to edges that share common vertices. The number of edges in the sample graph is $|E(G_s)| = \sum_{e \in E(G)}(1 - Y_e)$ which has mean

$$\mathbb{E}[|E(G_s)|] = \sum_{e \in E(G)} \mathbb{E}[(1 - Y_e)] = (1 - p^2)m$$

Therefore, dividing by $1 - p^2$ gives an unbiased estimator for the number of edges

$$\hat{m} = \frac{|E(G_s)|}{1 - p^2} \tag{3.9}$$

Next we calculate the variance of this estimator. The variance of the number of visible edges in the sample graph is

$$\mathrm{Var}\left(\sum_{e \in E}(1 - Y_e)\right) = \mathrm{Var}\left(\sum_{e \in E} Y_e\right) = \sum_{e \in E} \mathrm{Var}(Y_e) + \sum_{e \neq e'} \mathrm{Cov}(Y_e, Y_{e'})$$

Each term is examined separately. For the variance of each edge we have that $\mathrm{Var}(Y_e) = p^2(1 - p^2)$. For the covariance terms, if e and e' are disjoint edges the term is zero as they are independent random variables. However, in case they share a common vertex, for example $e = (u, v)$ and $e' = (v, w)$ then

$$\mathrm{Cov}(Y_e, Y_{e'}) = \mathbb{E}[Y_e Y_{e'}] - \mathbb{E}[Y_e]\,\mathbb{E}[Y_{e'}]$$

$$\begin{aligned}
&= \mathbb{E}[X_u X_v^2 X_w] - \mathbb{E}[Y_e]\,\mathbb{E}[Y_{e'}] \\
&= \mathbb{E}[X_u X_v X_w] - \mathbb{E}[Y_e]\,\mathbb{E}[Y_{e'}] \\
&= p^3 - p^2 p^2 \\
&= p^3(1-p)
\end{aligned}$$

To finish the calculation we must find how many dependent pairs of edges exist. Fix a vertex u and denote by $d_G(u)$ its degree in G. The number of correlated edge pairs with u as the common vertex is $d_G(u)(d_G(u)-1)$ (each pair counted twice as needed). Therefore

$$\begin{aligned}
\operatorname{Var}\left(\sum_{e\in E} Y_e\right) &= p^2(1-p^2)m + \sum_{u\in V} d_G(u)(d_G(u)-1)p^3(1-p) \\
&= p^2(1-p^2)m - 2p^3(1-p)m + \sum_{u\in V} d_G^2(u)p^3(1-p) \\
&= p^2(1-p)^2 m + p^3(1-p)\sum_{v\in V} d_G^2(v)
\end{aligned}$$

This is the exact expression for the variance. However, we will mostly need just an upper bound of it which we derive by using $d_G(v) \le d_G^2(v)$ as follows.

$$\begin{aligned}
\operatorname{Var}\left(\sum_{e\in E} Y_e\right) &\le \frac{1}{2}p^2(1-p)^2\sum_{v\in V} d_G^2(v) + p^3(1-p)\sum_{v\in V} d_G^2(v) \\
&= \frac{1}{2}p^2(1-p^2)\sum_{v\in V} d_G^2(v) \qquad (3.10)
\end{aligned}$$

Therefore, the variance of the estimator $\hat{m}$ is bounded from a term proportional to the sum of squares of degrees.

$$\operatorname{Var}(\hat{m}) \le \frac{p^2}{2(1-p^2)}\sum_{v\in V} d_G^2(v)$$

The ε-relative error we are interested is directly related to the variance via the Chebysev's inequality (2.1). An application of the inequality gives

$$\mathbb{P}(|\hat{m}-m| > \varepsilon m) \le \frac{\operatorname{Var}(\hat{m})}{\varepsilon^2 m^2} \le \frac{p^2}{2\varepsilon^2(1-p^2)}\frac{\sum_{v\in V} d_G^2(v)}{m^2}$$

From these calculations, we see that if p gets closer to 1 the bound gets worse as expected because if p is almost 1 then the sample should be very close to empty and the variance should be huge. Another conclusion is that it is the fraction $\sum_{v\in V} d_G^2(v)/m^2$ that determines the success of our estimator. If it is the case that

$$\frac{\sum_{v\in V} d_G^2(v)}{m^2} < \varepsilon^2\delta\frac{1-p^2}{p^2}$$

then we have an $\varepsilon-\delta$ estimator. There are cases where this is true, for example having a d-regular graph as the underlying graph. In this case $\frac{\sum_{v\in V} d_G^2(v)}{m^2} = \Theta(1/n)$

and thus we can have $\varepsilon - \delta$ estimation for very small values of ε and δ, even for $\varepsilon^2\delta = \Theta(1/n)$. Another example of practical interest because of their similarity with real world networks are power law graphs [B+16], that is, graphs with degree distribution $p_k = ck^{-\gamma}$ for γ typically in $(2, 3)$. It can be shown that for them the estimator's error decreases with n. On the other hand, the worst case is when the underlying graph is a star where the fraction $\frac{\sum_{v\in V} d_G^2(v)}{m^2} = \Theta(1)$. The fact that each vertex contributes to the variance proportionally to its squared degree will motivate us to believe that by choosing a few high degree vertices and querying for their true neighborhood will be enough to bring the estimation error down to δ, but this will be extensively discussed in the next chapter.

3.5.2 Upper Bound

Even if the underlying graph is a star, by drawing many samples from the distribution and using their union as more accurate sample, we can reduce the probability of error. As already noted, the union follows the same distribution with the only difference that the parameter p (the probability for each vertex to hide its neighborhood) is exponentially reduced. Thus p is replaced with p^N where N is the number of samples.

$$\begin{aligned}
\mathbb{P}(|\hat{m} - m| > \varepsilon m) &\leq \frac{p^{2N}}{2\varepsilon^2(1-p^{2N})} \frac{\sum_{v\in V} d_G^2(v)}{m^2} \\
&= \frac{2p^{2N}}{\varepsilon^2(1-p^{2N})} \frac{\sum_{v\in V} d_G^2(v)}{\left(\sum_{v\in V} d_G(v)\right)^2} \\
&\leq \frac{2p^{2N}}{\varepsilon^2(1-p)} \\
&< \delta
\end{aligned}$$

if $N > \frac{1}{2}\log\left(\frac{2}{\varepsilon^2\delta(1-p)}\right)$ samples are used. Note that the samples needed depend on ε, δ and p in a way that if p approaches 1 the quantity goes to infinity, as expected. Therefore, we proved the following.

Proposition 3.3. *The sample complexity of obtaining an (ε, δ)-estimator for the number of edges is $O\left(\log\left(\frac{1}{\varepsilon^2\delta}\right)\right)$ samples.*

3.5.3 Lower Bound

In order to avoid multiparametric bounds, we focus on the dependence on the parameter ε only (equivalently we can replace δ with $1/3$ and p with some other constant) to show that the lower bound mathces the upper bound of $\log(1/\varepsilon)$ given previously. We will reduce the estimating problem to a hypothesis testing problem as demonstrated in Section 2.5. This is essentially the same procedure as in the proof of Proposition 2.5, with the difference that here we are not dealing with the minimax error but with the ε-relative error. The same reduction can be found in other works too, such as [Hub17].

Proposition 3.4. *Every estimator $\hat{m}$ that satisfies $\mathbb{P}_{\mathbf{X}\sim P_G^N}(|\hat{m}(\mathbf{X}) - m_G| > \varepsilon m_G) < 1/3$ for every graph G needs $N = \Omega(\log(1/\varepsilon))$ samples.*

Proof. Consider the following hypothesis testing that is very similar to the one presented in the last lower bound construction. The graph for the null hypothesis H_0 has vertex set $\{u_1, \ldots, u_n\} \cup \{v_1, \ldots, v_n\}$ and edge set $\{(u_i, v_i)\}_{i=1}^n$, that is, the graph is a matching with n edges.

The alternative hypothesis H_1 is a composite hypothesis, where a graph is selected uniformly at random from the family of all graphs that are defined like the graph of the null hypothesis with the only difference that they have $n(1-\varepsilon)/(1+\varepsilon)$ edges instead of n.

Suppose there exists an estimator $\hat{m}$ that satisfies $\mathbb{P}_{\mathbf{X}\sim P_G^N}(|\hat{m}(\mathbf{X}) - m_G| > \varepsilon m_G) < 1/3$ for every graph G, with $N = o(\log(1/\varepsilon))$ samples, to derive a contradiction. We can use this estimator to solve the hypothesis testing with probability of error less than $1/3$. Indeed, define the testing function Ψ to be

$$\Psi(\mathbf{X}) = \begin{cases} 0, & \hat{m}(\mathbf{X}) \geq n(1-\varepsilon) \\ 1, & \hat{m}(\mathbf{X}) < n(1-\varepsilon) \end{cases}$$

It can be easily seen that the error of this estimator (were we will denote by V the random variable that determines which hypothesis is selected), $\mathbb{P}(\Psi \neq V) = \frac{1}{2}\mathbb{P}_{H_0}(\Psi \neq 0) + \frac{1}{2}\mathbb{P}_{H_1}(\Psi \neq 1)$ is less than $1/3$ because each term is less than $1/3$. For example the first term is

$$\mathop{\mathbb{P}}_{H_0}(\Psi \neq 0) = \mathbb{P}(\hat{m}(\mathbf{X}) < n(1-\varepsilon)) = \mathbb{P}(\hat{m}(\mathbf{X}) < m(1-\varepsilon)) < 1/3$$

If we choose $\varepsilon = 1/(2n-1)$, the graphs of the null and alternative hypotheses have n and $n-1$ edges respectively. This is the hypothesis testing examined in the proof of Theorem 3.2, where we showed that $\Omega(\log n)$ samples are required for this hypothesis testing problem to be solved with probability of error less than $1/3$. However, we assumed before that the estimator we used works with $o(\log \varepsilon^{-1})$ samples which is $o(\log n)$ if $\varepsilon = 1/(2n-1)$. We get a contradiction. ■

The two propositions combined determine the sample complexity of approximate edge estimation.

Theorem 3.3. *There exists an estimator $\hat{m}$ such that for every $\varepsilon, \delta \in (0, 1]$ and graph G with m edges, given $N = \Theta(\log(1/\varepsilon^2\delta))$ i.i.d. samples $X_1, \ldots, X_N \sim \mathcal{P}_G$ satisfies $\mathbb{P}_{\mathbf{X}\sim\mathcal{P}_G^N}(|\hat{m} - m| > leq\varepsilon m) < \delta$. In addition, if $N = o(\log \varepsilon^{-1})$ then for every estimator $\hat{m}$ there exists a graph G such that $\mathbb{P}_{\mathbf{X}\sim\mathcal{P}_G^N}(|\hat{m} - m| > \varepsilon m) \geq 1/3$.*

3.6 Approximate Estimation of Triangles

We extend the previous efforts to obtain the same guaranties for the case of estimating triangles. We will essentially show that the same approach gives similar results and the difference is a slightly higher level of complexity regarding the calculations involved. In particular, as it will be seen below, the expression for the variance of the estimation here is a direct generalization of the corresponding expression for the triangles.

3.6.1 Mean and Variance

Let $X_v \sim Bernoulli(p)$ be the random trials associated with each vertex $v \in V$ and T the set of all triangles of the underlying graph G. Also, let G_s be a sample graph and $T(G_s)$ the set of its triangles. For each triangle $t \in T$ define the indicator random variable that equals to one if and only if the triangle is hidden from the sample graph.

$$Y_t = \begin{cases} 1, & t \notin T(G_s) \\ 0, & t \in T(G_s) \end{cases}$$

We calculate the expected value of the total number of triangles in the sample graph $|T(G_s)| = \sum_{t \in T}(1 - Y_t)$. To do so, note that the probability for a triangle (u, v, w) to be visible is

$$\begin{aligned} \mathbb{P}(X_u + X_v + X_w \geq 2) &= \mathbb{P}(X_u + X_v + X_w \geq 2) + \mathbb{P}(X_u + X_v + X_w \geq 3) \\ &= 3p^2(1-p) + p^3 = p^2(3 - 2p) \end{aligned}$$

Therefore the expectation we are interested in is

$$\mathbb{E}\left[\sum_{t \in T}(1 - Y_t)\right] = |T|(1 - p^2(3 - 2p)) = |T|(1 - 3p^2 + 2p^3)$$

Dividing by the factor $1 - 3p^2 + 2p^3$ gives an unbiased estimator for the number of triangles

$$\hat{T} = \frac{|T(G_s)|}{1 - 3p^2 + 2p^3}$$

Next we calculate the variance of the number of triangles in the underlying graph

$$\mathrm{Var}\left(\sum_{t \in T}(1 - Y_t)\right) = \mathrm{Var}\left(\sum_{t \in T} Y_t\right) = \sum_{t \in T} \mathrm{Var}(Y_t) + \sum_{t \neq t'} \mathrm{Cov}(Y_t, Y_{t'})$$

For the first terms we have that

$$\begin{aligned} \mathrm{Var}(Y_t) &= \mathbb{E}[Y_t^2] - \mathbb{E}[Y_t]^2 \\ &= \mathbb{E}[Y_t] - \mathbb{E}[Y_t]^2 \\ &= \mathbb{E}[Y_t](1 - \mathbb{E}[Y_t]) \\ &= 3p^2 - 2p^3 - 9p^4 + 12p^5 - 4p^6 \end{aligned}$$

For the covariance terms, we have that the following depending on the number of shared vertices.

1. If the two triangles are disjoint, then the corresponding covariance term is zero

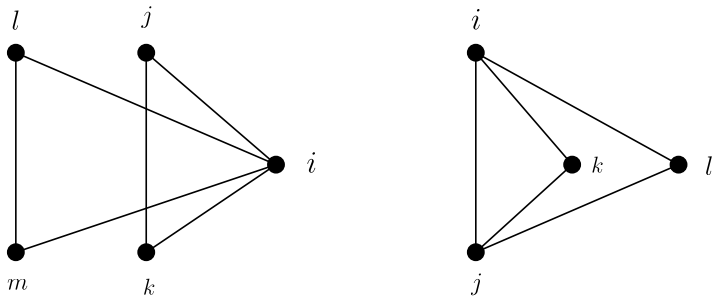

Figure 3.2. Pair of triangles with one shared vertex and pair with one shared edge.

2. It t and t' share a common vertex (it does not matter which of the three) according to Figure 3.2, then $\mathrm{Cov}(Y_{ijk}, Y_{ilm}) = \mathbb{E}[Y_{ijk}Y_{ilm}] - \mathbb{E}[Y_{ijk}]E[Y_{ilm}] = \mathbb{E}[Y_{ijk}Y_{ilm}] - (p^2(3-2p))^2$ where

$$\begin{aligned}\mathbb{E}[Y_{ijk}Y_{ilm}] &= \mathbb{P}(Y_{ijk}=1, Y_{ilm}=1)\\ &= \mathbb{P}(Y_{ijk}=1, Y_{ilm}=1 \mid X_i=1)p + \mathbb{P}(Y_{ijk}=1, Y_{ilm}=1 \mid X_i=0)(1-p)\\ &= (p^2+2p(1-p))^2p + p^4(1-p)\\ &= (4-3p)p^3\end{aligned}$$

   Therefore, $\mathrm{Cov}(Y_{ijk}, Y_{ilm}) = 4p^3(1-p)^3$ in this case.

3. If t and t' share two common vertices we have that

$$\begin{aligned}\mathbb{E}[Y_{ijk}Y_{ilm}] &= \mathbb{P}(Y_{ijk}Y_{ilm}=1 \mid X_i=1, X_j=1)p^2\\ &\quad + 2\,\mathbb{P}(Y_{ijk}Y_{ilm}=1 \mid X_i=0, X_j=1)p(1-p)\\ &\quad + \mathbb{P}(Y_{ijk}Y_{ilm}=1 \mid X_i=0, X_j=0)p(1-p)^2\\ &= p^2 + 2p^2p(1-p)\\ &= p^2(1-p)^2(1+4p-4p^2)\end{aligned}$$

   Therefore, $\mathrm{Cov}(Y_{ijk}, Y_{ilm}) = p^2(1-p)^2(1+4p-4p^2)$ in this case.

The goal is to find a concrete expression for the variance, a formula that would be analogous to the case of edge estimation. In edge estimation we had the sum of squares of degrees. Here, instead of degrees we will have the number of triangles incident to each vertex, which are, in a way, "triangle degrees".

To begin, observe that the covariance term for triangles t_1, t_2 for both of the cases presented above can be written as

$$\mathrm{Cov}(Y_{t_1}, Y_{t_2}) = \begin{cases} 4p^3(1-p)^3, & \text{one shared vertex} \\ 4p^3(1-p)^3 + p^2(1-p)^2, & \text{two shared vertices} \end{cases}$$

This means that the term for the case of two shared vertices is equal to the term for the case of one shared vertex plus another term. Using the adjacency matrix A

of the underlying graph, we have the following for the sum of all covariance terms

$$\sum_{k\neq l}\mathrm{Cov}(Y_{t_k},Y_{t_l}) = \sum_{v\in V} 4p^3(1-p)^3\frac{A^3[v,v]}{2}\left(\frac{A^3[v,v]}{2}-1\right) + \sum_{(u,v)\in E} A^2[u,v]\left(A^2[u,v]-1\right)\left(p^2(1-p)^2-4p^3(1-p)^3\right)$$

To provide an explanation for the above expression, fix a vertex v. The value of $A^3[v,v]$ is equal to twice the number of triangles incident to that vertex. For each pair of such triangles $4p^3(1-p)^3$ must be added. However, there may exist triangles that share two vertices. For those pairs, we must add an extra $p^2(1-p)^2$. The number of those triangles are determined by A^2. The reason $4p^3(1-p)^3$ is subtracted is because the triangles that share a common edge, also share a common vertex and thus $4p^3(1-p)^3$ has already been added twice (one time for each endpoint of the shared edge).

By some extra algebraic manipulations, we obtain a final expression of the variance.

$$\mathrm{Var}\left(\sum_{t\in T}Y_t\right) = |T|4p^3(1-p)^3 + p^3(1-p)^3\sum_{v\in V}(A^3[v,v])^2 + p^2(1-p)^2(1-2p)^2\sum_{(u,v)\in E}(A^2[u,v])^2$$

Denote by $\lambda(v)$ the number of triangles of G that are incident to vertex v and by $\lambda(u,v)$ the number of triangles of G that include the edge $(u,v)\in E$. We will use these instead of the adjacency matrix and also we will use the fact that $|T| = \sum_{v\in V}\lambda(v)/3$ as well as $\lambda(v)\leq\lambda^2(v)$ to obtain a simpler upper bound of the variance.

$$\begin{aligned}\mathrm{Var}\left(\sum_{t\in T}Y_t\right) &= \frac{4}{3}p^3(1-p)^3\sum_{v\in V}\lambda(v)\\ &\quad + 4p^3(1-p)^3\sum_{v\in V}\lambda^2(v) + p^2(1-p)^2(1-2p)^2\sum_{(u,v)\in E}\lambda^2(u,v)\\ &\leq \frac{4}{3}p^3(1-p)^3\sum_{v\in V}\lambda^2(v) + 4p^3(1-p)^3\sum_{v\in V}\lambda^2(v)\\ &\quad + p^2(1-p)^2(1-2p)^2\sum_{(u,v)\in E}\lambda^2(u,v)\\ &\leq 6p^3(1-p)^3\sum_{v\in V}\lambda^2(v) + p^2(1-p)^2(1-2p)^2\sum_{(u,v)\in E}\lambda^2(u,v)\end{aligned}$$

Finally we can express the ε-relative error using Chebysev's inequality (2.1).

$$\mathbb{P}(|\hat{T}-T|>\varepsilon T)\leq\frac{\mathrm{Var}(\hat{T})}{\varepsilon^2T^2}\leq\frac{c_1(p)\sum_{v\in V}\lambda^2(v)+c_2(p)\sum_{(u,v)\in E}\lambda^2(u,v)}{\varepsilon^2T^2}$$

where $c_1(p) = 6p^3(1-p)^3/(1-3p^2+2p^3)^2$ and $c_2(p) = p^2(1-p)^2(1-2p)^2/(1-3p^2+2p^3)^2$ are constants depending only on p.

The expression for the variance shows that the crucial fractions from which the estimation's error depends are $\sum_{v\in V}\lambda^2(v)/T^2$ and $\sum_{(u,v)\in E}\lambda^2(u,v)/T^2$. Again,

for certain underlying graphs such as regular graphs, these quantities are $o(1)$ and thus the estimation has error that naturally goes to zero. However, the worst case here would be to have a graph with vertex set $\{v\} \cup \{u_1, \ldots, u_{2n}\}$ and edge set $\{(v, u_i)\}_{i=1}^{2n} \cup \{(u_i, u_{2n-i+1})\}_{i=1}^{n}$, that is, n triangles sharing one central vertex v. In this case the upper bound becomes constant. A query strategy discussed in the following chapter will focus on eliminating these vertices of high "triangle degree" from the variance.

3.6.2 Upper Bound

Similarly to the edge estimation case, if N samples are available, we can use their union to improve the error of estimation. The reason is that if we use the union, the parameter p will be replaced by p^N, and thus the variance will be reduced exponentially in the number of samples, as we show below. Using the fact that $\sum_{(u,v)\in E} \lambda^2(u,v) \leq \sum_{v\in V} \lambda^2(v)$ and that $c_1(p^N) \leq 6p^{3N}/(1-3p^2+2p^3)^2$, $c_2(p^N) \leq p^{2N}/(1-3p^2+2p^3)^2$ we have

$$\begin{aligned}
\mathbb{P}(|\hat{T} - T| > \varepsilon T) &\leq \frac{c_1(p^N)\sum_{v\in V}\lambda^2(v) + c_2(p^N)\sum_{(u,v)\in E}\lambda^2(u,v)}{\varepsilon^2 T^2} \\
&\leq \frac{[c_1(p^N) + c_2(p^N)]\sum_{v\in V}\lambda^2(v)}{\varepsilon^2 T^2} \\
&= \frac{6p^{3N} + p^{2N}}{(1-3p^2+2p^3)^2} \frac{9\sum_{v\in V}\lambda^2(v)}{\varepsilon^2\left(\sum_{v\in V}\lambda(v)\right)^2} \\
&\leq \frac{9(6p^{3N} + p^{2N})}{(1-3p^2+2p^3)^2\varepsilon^2} \\
&\sim \frac{9p^{2N}}{(1-3p^2+2p^3)^2\varepsilon^2}
\end{aligned}$$

By demanding that this expression is less than δ and solving the inequality, we find that the number of samples N must be at least logarithmic in $\varepsilon^{-2}\delta^{-1}$. We state this conclusion below.

Proposition 3.5. *The sample complexity of obtaining an (ε, δ)-estimator for the number of triangles is $O\left(\log\left(\frac{1}{\varepsilon^2\delta}\right)\right)$.*

3.6.3 Lower Bound

To avoid repeating the same proofs, we state that the same lower bound of the case of edge estimation holds true for triangles as it can be shown in exactly the same way.

Proposition 3.6. *Every estimator $\hat{T}$ that satisfies $\mathbb{P}_{\mathbf{X}\sim P_G^N}(|\hat{T}(\mathbf{X}) - T| > \varepsilon T) < 1/3$ for every graph G needs $N = \Omega(\log(1/\varepsilon))$ samples.*

Theorem 3.4. *There exists an estimator $\hat{t}$ such that for every $\varepsilon, \delta \in (0, 1]$ and graph G with t triangles, given $N = \Theta(\log(1/\varepsilon^2\delta))$ i.i.d. samples $X_1, \ldots, X_N \sim \mathcal{P}_G$ satisfies $\mathbb{P}_{\mathbf{X}\sim\mathcal{P}_G^N}(|\hat{t} - t| > \varepsilon t) < \delta$. In addition, if $N = o(\log \varepsilon^{-1})$ then for every estimator $\hat{t}$ there exists a graph G such that $\mathbb{P}_{\mathbf{X}\sim\mathcal{P}_G^N}(|\hat{t} - t| > \varepsilon t) \geq 1/3$.*

Remark 3.5. *These conclusions should also hold for other subgraphs such as circles of length more than three. The expression of the variance should be easily generalized for these cases and the same logarithmic bound should hold true.*

Chapter 4

Estimation Using Two Samples and Some Queries

In this chapter we allow oracle access to the underlying graph in addition to the samples. More specifically, the learning algorithms may perform queries to learn the true neighborhood of vertices which they choose. Based on the analysis of the variance done in Chapter 3 we specify which are the critical vertices for the properties examined and we design query strategies to eliminate their contribution to the estimation's error. The resulting estimators require two full samples of the underlying graph and perform a number of queries that depends only on the parameters of accuracy and not the size of the graph. The result is summarized below.

Theorem 4.1. *(Informal) There exists an estimator $\hat{m}$ which uses two samples and $k = \Theta(\varepsilon^{-2}\delta^{-1})$ queries and satisfies $\mathbb{P}(|\hat{m} - m| > \varepsilon m) \leq \delta$.*

For the triangles, we get a somewhat weaker guarantee.

Theorem 4.2. *(Informal) There exists an estimator $\hat{T}$ which uses two samples and $k = \Theta(\varepsilon^{-2}\delta^{-1})$ queries and satisfies $\mathbb{P}(|\hat{T} - T| > \varepsilon W) \leq \delta$, where W is the total number of wedges in graph G.*

4.1 Query Strategy

Based on the observations and conclusions of the previous chapter, we start designing the strategy for selecting which vertices to query. As a warm-up we first study the behavior of the ε-relative error of the edge estimator $\hat{m} = |E(G_s)|/(1-p^2)$, which is restated here.

$$\mathbb{P}(|\hat{m} - m| > \varepsilon m) \leq \frac{p^2}{2\varepsilon^2(1-p^2)} \frac{\sum_{v \in V} d_G^2(v)}{m^2}$$

in order to find out which are the cases where this error is naturally decreasing with the size of the graph (and thus the need for queries is obviated) and how do the graphs that do not belong in these cases look like. Some observations on this matter are the following.

Claim 4.1. *If $m = \omega(n)$, the estimator's relative error (without queries) asymptotically vanishes.*

Proof.

$$\sum_{u\in V} d_G^2(u) = \sum_{(u,v)\in E} (d_G(u) + d_G(v)) \leq 2mn$$
$$\frac{\sum_u d_G^2(u)}{m^2} \leq 2\frac{n}{m} \xrightarrow{m=\omega(n)} 0$$

■

This shows that if it is the case that $m = \omega(n)$, then no queries are needed to correct the estimator, as it already has zero error asympotically. We will now show that in any other case, that is, if the error is initially high, then only a constant number of vertices is responsible for that.

Claim 4.2. *Assume that* $\sum_{u\in V} d_G^2(u)/m^2 = \Theta(1)$. *Then, there exist* $k = \Theta(1)$ *vertices of degree* $\Theta(m)$.

Proof. Let us assume that all degrees are $o(m)$ to derive a contradiction.

$$\frac{\sum_u d_G^2(u)}{m^2} = \frac{\sum_{(u,v)}(d_G(u) + d_G(v))}{m^2} \leq \frac{o(m)m}{m^2}$$

which means that $\sum_{u\in V} d_G^2(u)/m^2 \to 0$. Therefore, there must be a vertex with degree $\Theta(m)$. Additionally, the number of vertices of such degree cannot be more than constant. ■

This suggests that every hard instance must look like a small collection of stars, which is what was intuitively expected. Using the fact that each vertex we query "disappears" from the variance (which we will show through the analysis later), we get that the number of queries needed to reduce the error of estimation is independent of the size of the graph.

Based on the previous discussion, the ideal scenario would be to query for the vertices of highest degree in the underlying graph. However, the underlying graph is unknown and thus this is an unrealistic set up. Nevertheless, we are going to analyze it, because, we will later show that using the same strategy applied on the sample graph, the results are similar (because we will show that the degree ranking does not fundamentally change in the sample graph).

Notation. For every set S of vertices, let $G[S]$ denote the subgraph that is induced from the set S, that is, the subgraph which has S as its vertex set and $\{(u, v) \in E(G) \mid u \in S, v \in S\}$ as its edge set.

Let Q be a set of vertices to be queried, which we will call *query set*. For the reasons explained, for now we will assume that the query set is deterministically specified. Define the estimator

$$\hat{m} = \sum_{e=(u,v)\in E} \mathbf{1}(u \in Q \vee v \in Q) + \frac{1}{1-p^2} \sum_{e\in E(G[V\setminus Q])} \mathbf{1}(e \in E(G_1))$$

This estimator is unbiased, as we have that (note that Q is not random)

$$\mathbb{E}[\hat{m}] = \sum_{(u,v)\in E} \mathbf{1}(u \in Q \vee v \in Q) + \sum_{(u,v)\in E} \mathbf{1}(u \notin Q, v \notin Q)\frac{1-p^2}{1-p^2} = m \qquad (4.1)$$

Now we calculate its variance to show that the vertices from Q "disappear" from the formula we gave in the previous chapter. Since the first term is deterministic, its variance is zero, thus we examine the other term

$$\begin{aligned}\operatorname{Var}\left[\sum_{e\in E(G[V\setminus Q])}(1-Y_e)\right] &= \operatorname{Var}\left[\sum_{e\in E(G[V\setminus Q])}Y_e\right]\\ &\overset{(3.10)}{=} \frac{1}{2}p^2(1-p^2)\sum_{v\in V\setminus Q} d^2_{G[V\setminus Q]}(v)\\ &\leq \frac{1}{2}p^2(1-p^2)\sum_{v\in V\setminus Q} d^2_G(v) \qquad (4.2)\end{aligned}$$

This shows why the queried vertices disappear from the variance. We conclude by bounding the number of queries needed to bring this variance divided by $\varepsilon^2 m^2$ (that is, the estimation's probability of ε-relative error) down to δ. Claim 4.2 suggests that this number is constant. A more precise bound depending on the parameters ε, δ is $1/\varepsilon^2\delta$, which is stated as Lemma 4.3.

4.2 Analysis of the Estimator

Having analyzed the "optimal" query strategy, where the highest degree vertices of the underlying graph are queried, we move to the more realistic set up where we need to approximately find those vertices from the sample graph. The strategy here is to apply the same selection rule to the sample graph. It remains to prove that this is a good approximation.

We first give a complete description of the estimator. Suppose we are given two independent samples $G_1, G_2 \sim \mathcal{P}_G$. The first sample will be used to determine the query set and the second to calculate the estimation. The reason we require different samples is that technical difficulties arise when the query set is statistically correlated with the sample graph, which are discussed at the end of the section. From the first sample G_1, a query set $Q(G_1)$ of the k highest degree vertices is determined. After the set Q is queried, all the neighborhoods $\{\Gamma_G(u) \mid u \in Q\}$ become known (we will use subscripts to be clear about which graph we consider each time). Then, the second sample is used along with the information from the queries to output the estimation which is calculated by counting the edges that are adjacent to Q (first term below) and the edges of the second sample graph after the removal of the set Q from it (second term)

$$\hat{m} = \sum_{(u,v)\in E(G)} \mathbf{1}(u \in Q(G_1) \vee v \in Q(G_1)) + \frac{1}{1-p^2}\sum_{e\in G[V\setminus Q]} \mathbf{1}(e \in E(G_2)) \qquad (4.3)$$

Lemma 4.1. *For the expected value and variance of the estimator we have the following*

$$\mathop{\mathbb{E}}_{G_1,G_2\sim\mathcal{P}_G}[\hat{m}] = m$$

$$\operatorname*{Var}_{G_1,G_2\sim\mathcal{P}_G}[\hat{m}] \leq c(p) \operatorname*{\mathbb{E}}_{G_1\sim\mathcal{P}_G}\left[\sum_{u\in V\setminus Q(G_1)} d_G^2(u)\right]$$

where $c(p)$ depends only on the parameter p of each vertex to hide.

Proof. We use the law of total variance

$$\begin{aligned}
\operatorname*{Var}_{G_1,G_2\sim\mathcal{P}_G}(\hat{m}) &= \operatorname*{\mathbb{E}}_{G_1\sim\mathcal{P}_G}[\operatorname*{Var}_{G_2\sim\mathcal{P}_G}(\hat{m})] + \operatorname*{Var}_{G_1\sim\mathcal{P}_G}[\operatorname*{\mathbb{E}}_{G_2\sim\mathcal{P}_G}[\hat{m}]] \\
&= \operatorname*{\mathbb{E}}_{G_1\sim\mathcal{P}_G}[\operatorname*{Var}_{G_2\sim\mathcal{P}_G}(\hat{m})] + \operatorname*{Var}_{G_1\sim\mathcal{P}_G}[m] \\
&= \operatorname*{\mathbb{E}}_{G_1\sim\mathcal{P}_G}[\operatorname*{Var}_{G_2\sim\mathcal{P}_G}(\hat{m})] + 0 \\
&\leq c(p) \operatorname*{\mathbb{E}}_{G_1\sim\mathcal{P}_G}\left[\sum_{u\in V\setminus Q(G_1)} d_G^2(u)\right]
\end{aligned}$$

The last inequality follows from the calculation of the estimator's variance when the query set is deterministic (4.2). ■

Lemma 4.2. *Let Q^* be the set of the $k < n$ vertices of highest degree in the underlying graph G and let $Q(G_1)$ be the same set for the sample graph $G_1 \sim \mathcal{P}_G$. Under Assumption 1 (specified at the end of the proof), there exists a constant $a = a(p)$ that depends only on the parameter p of each vertex to hide, such that*

$$\operatorname*{\mathbb{E}}_{G_1\sim\mathcal{P}_G}\left[\sum_{u\in V\setminus Q(G_1)} d_G^2(u)\right] \leq a \sum_{u\in V\setminus Q^*} d_G^2(u)$$

Proof. Denote by $\hat{U}(G_1)$ the set of the $l = n-k$ lowest degree vertices in G_1 and by U^* the same set for G. The lemma is restated as follows. We want to find a constant a such that

$$\operatorname*{\mathbb{E}}_{G_1\sim\mathcal{P}_G}\left[\sum_{u\in \hat{U}(G_1)} d_G^2(u)\right] \leq a \sum_{u\in U^*} d_G^2(u)$$

An assumption is needed about the degrees because if the graph has only low degree vertices, their ranking can be completely changed in the sample graph. Let $\hat{u}_i$ denote the vertex with the i-th smallest degree in G_1 and let u_i^* denote the same thing for G. Then, the left hand side of the previous inequality is

$$\begin{aligned}
\operatorname*{\mathbb{E}}_{G_1\sim\mathcal{P}_G}\left[\sum_{u\in \hat{U}(G_1)} d_G^2(u)\right] &= \operatorname*{\mathbb{E}}_{G_1\sim\mathcal{P}_G}[d_G^2(\hat{u}_1)] + \cdots + d_G^2(\hat{u}_l)] \\
&= \operatorname*{\mathbb{E}}_{G_1\sim\mathcal{P}_G}[d_G^2(\hat{u}_1)] + \cdots + \operatorname*{\mathbb{E}}_{G_1\sim\mathcal{P}_G}[d_G^2(\hat{u}_l)]
\end{aligned}$$

The main idea is that we would like to prove for each one of the terms the desired inequality, namely that $\mathbb{E}_{G_1}[d_G^2(\hat{u}_i)] \leq (2/(1-p))^2 d_G^2(u_i^*)$. To do so, we examine an arbitrarily selected vertex v. If $d_G(v) > 2d_G(u_i^*)/(1-p)$ it is almost impossible that

this vertex will be chosen as the the i-th smallest in the sample because even if it is hidden, its degree in G_1 will be concentrated around its expectation $\mathbb{E}[d_{G_1}(v)] > 2d_G(u_i^*) > d_G(u_i^*)$ (this means that there will be at least i vertices with smaller degrees in the sample and thus $\hat{u}_i \neq v$). This is formalized using Chernoff bounds (Theorem 2.3): $\mathbb{P}(\hat{u}_i = v) \leq \mathbb{P}(d_{G_1}(v) < \mathbb{E}[d_{G_1}(v)]/2) \leq \exp(-\mathbb{E}[d_{G_1}(v)]/8) = \exp(-(1-p)d_G(v)/8)$. For the vertices of high degree the bound works, but for the vertices of low degree we have to work independently. Thus we do the following

$$\begin{aligned}
\mathop{\mathbb{E}}_{G_1\sim\mathcal{P}_G}\left[\sum_{u\in\hat{U}(G_1)} d_G^2(u)\right] &= \sum_{i=1}^{l} \mathop{\mathbb{E}}_{G_1\sim\mathcal{P}_G}[d_G^2(\hat{u}_i)] \\
&= \sum_{i=1}^{l}\sum_{v\in V} d_G^2(v) \mathop{\mathbb{P}}_{G_1\sim\mathcal{P}_G}(\hat{v}_i = v) \\
&= \sum_{i=1}^{l}\Bigg\{ \sum_{v\,:\,d_G(v)\leq\frac{2d_G(u_i^*)}{1-p}} d_G^2(v) \mathop{\mathbb{P}}_{G_1\sim\mathcal{P}_G}(\hat{u}_i = v) \\
&+ \sum_{v\,:\,\frac{2d_G(u_i^*)}{1-p}<d_G(v)\leq\frac{24\log n}{1-p}} d_G^2(v) \mathop{\mathbb{P}}_{G_1\sim\mathcal{P}_G}(\hat{u}_i = v) \\
&+ \sum_{v\,:\,\frac{24\log n}{1-p}<d_G(v)} d_G^2(v) \mathop{\mathbb{P}}_{G_1\sim\mathcal{P}_G}(\hat{u}_i = v)\Bigg\} \\
&= A + B + C
\end{aligned}$$

We examine each term bellow. For the first one we have that

$$A = \sum_{i=1}^{l} \sum_{v\,:\,d_G(v)\leq\frac{2d_G(u_i^*)}{1-p}} d_G^2(v) \mathop{\mathbb{P}}_{G_1\sim\mathcal{P}_G}(\hat{u}_i = v) \leq \frac{4}{(1-p)^2}\sum_{i=1}^{l} d_G^2(u_i^*)$$

For the second term we have

$$\begin{aligned}
B &= \sum_{i=1}^{l} \sum_{v\,:\,\frac{2d_G(u_i^*)}{1-p}<d_G(v)\leq\frac{24\log n}{1-p}} d_G^2(v) \mathop{\mathbb{P}}_{G_1\sim\mathcal{P}_G}(\hat{u}_i = v) \\
&\leq \sum_{i=1}^{l} \sum_{v\,:\,d_G(v)\leq\frac{24\log n}{1-p}} d_G^2(v) \mathop{\mathbb{P}}_{G_1\sim\mathcal{P}_G}(\hat{u}_i = v) \\
&= \sum_{v\,:\,d_G(v)\leq\frac{24\log n}{1-p}} d_G^2(v) \sum_{i=1}^{l} \mathop{\mathbb{P}}_{G_1\sim\mathcal{P}_G}(\hat{u}_i = v) \\
&\leq \sum_{v\,:\,d_G(v)\leq\frac{24\log n}{1-p}} d_G^2(v)
\end{aligned}$$

For the last term we have

$$\begin{aligned}
C &= \sum_{i=1}^{l} \sum_{v\,:\,\frac{24\log n}{1-p} < d_G(v)} d_G^2(v) \mathop{\mathbb{P}}_{G_1\sim\mathcal{P}_G}(\hat{u}_i = v) \\
&\leq \sum_{i=1}^{l} \sum_{v\,:\,\frac{24\log n}{1-p} < d_G(v)} d_G^2(v) \mathop{\mathbb{P}}_{G_1\sim\mathcal{P}_G}\left(d_{G_1}(v) < \frac{\mathbb{E}[d_{G_1}(v)]}{2}\right) \\
&\leq \sum_{i=1}^{l} \sum_{v\,:\,\frac{24\log n}{1-p} < d_G(v)} d_G^2(v) \exp\left(-(1-p)\frac{d_G(v)}{8}\right) \\
&\leq \sum_{i=1}^{l} \sum_{v\,:\,\frac{24\log n}{1-p} < d_G(v)} d_G^2(v) \exp\left(-\frac{1-p}{8}\frac{24\log n}{1-p}\right) \\
&\leq n^3 \exp(-3\log n) \\
&= 1
\end{aligned}$$

Where we used the multiplicative Chernoff bound for the second inequality. Therefore, we can bound the fraction as follows:

$$\frac{\mathbb{E}_{G_1\sim\mathcal{P}_G}\left[\sum_{u\in\hat{U}(G_1)} d_G^2(u)\right]}{\sum_{u\in U^*} d_G^2(u)} \leq \frac{4}{(1-p)^2} + \frac{1}{\sum_{u\in U^*} d_G^2(u)} + \frac{1}{\sum_{u\in U^*} d_G^2(u)} \sum_{v\,:\,d_G(v)\leq\frac{24\log n}{1-p}} d_G^2(v)$$

Assumption 1. G has more than k vertices of degree higher than $\frac{24\log n}{1-p}$. Therefore, the second term is less than 1 and the final term also less than 1 because of our assumption. ■

The above Lemma states that the error of the estimator is an approximation of the error under the "optimal" query strategy. It is now time to determine how many queries are required by the optimal query strategy.

Lemma 4.3. *Let Q^* be the set of the k vertices of highest degrees in the underlying graph G. For $k = \Theta(\varepsilon^{-2}\delta^{-1})$ we have*

$$\frac{\sum_{u\in V\setminus Q^*} d_G^2(u)}{m^2\varepsilon^2} \leq \delta$$

Proof. Let k be the size of Q^*. Also let $d_1 \geq d_2 \geq \cdots \geq d_k \geq \cdots \geq d_n$ the degree sequence of the underlying graph G. Define $c_i = d_i/m$.

$$\begin{aligned}
f(G) = \frac{\sum_{u\in V\setminus Q^*} d_G^2(u)}{m^2\varepsilon^2} &= \frac{1}{\varepsilon^2}\left(\frac{d_{k+1}^2}{m^2} + \cdots + \frac{d_n^2}{m^2}\right) \\
&= \frac{1}{\varepsilon^2}(c_{k+1}^2 + \cdots + c_n^2)
\end{aligned}$$

Now we may think of an adversary, who knows k, picking the worst case degrees for the underlying graph, equivalently we want to maximize this function under the

constraint $\sum_{i=1}^{n} c_i = 2$. We make two observations about the form of the optimal solution. The first one is that the optimal choice of the c_i's will have $c_1 = \cdots = c_k = c_{k+1}$ because that way the mass of the first c_i's that do not appear in our sum will be minimized. The second observation is that the optimal solution can contain zeros. More typically, we have to solve the following constrained optimization problem (k and n are fixed)

$$\begin{aligned} \max_{y,x_{k+2},\ldots,x_n} & f(y, x_{k+2}, \ldots, x_n) \\ g(y, x_{k+2}, \ldots, x_n) &= 0 \\ x_{k+2} &\leq y \\ x_{k+3} &\leq y \\ &\vdots \\ x_n &\leq y \end{aligned}$$

Where $f(y, x_{k+2}, \ldots, x_n) = y^2 + x_{k+2}^2 + \cdots + x_n^2$ and $g(y, x_{k+2}, \ldots, x_n) = (k+1)y + x_{k+2} + \cdots + x_n - 2$. The optimal solution is described by the KKT conditions. Yet, our inequality constraints are very easy and thus we can just examine two cases: the first is when the inequalities are inactive (not tight) and the second is when some or all of the inequalities are tight.

Case 1 (inactive inequalities). We solve the corresponding inequality-free problem with Lagrange multipliers and then just verify that the inequalities hold.

$$\begin{aligned} \nabla f(y, x_{k+2}, \ldots, x_n) &= \lambda \nabla g(y, x_{k+2}, \ldots, x_n) \\ \Rightarrow 2y = \lambda(k+1), 2x_{k+2} &= \lambda, \ldots, 2x_n = \lambda \end{aligned}$$

Using these along with the equation $(k+1)y + x_{k+2} + \cdots + x_n = 2$, we solve for $\lambda = 4/[(k+1)^2 + (n-k-1)]$ and plug it in to find $y = 2(k+1)/[(k+1)^2 + (n-k-1)]$ and $x_{k+2} = \ldots = x_n = 2/[(k+1)^2 + (n-k-1)]$. The corresponding value of f on this point is

$$f(y^*, x_{k+2}^*, \ldots, x_n^*) = \frac{4}{(k+1)^2 + (n-k-1)}$$

Case 2 (tight inequalities). Without loss of generality we can assume that the first μ inequalities are tight, that is our cost function becomes $f(y, x_{k+2+\mu}, \ldots, x_n) = (\mu+1)y^2 + x_{k+2+\mu}^2 + \cdots + x_n^2$ and the constraint $g(y, x_{k+2+\mu}, \ldots, x_n) = (k+\mu+1)y + x_{k+2+\mu} + \cdots + x_n - 2$. We solve this optimization problem with Lagrange multipliers.

$$\begin{aligned} \nabla f(y, x_{k+2+\mu}, \ldots, x_n) &= \lambda \nabla g(y, x_{k+2+\mu}, \ldots, x_n) \\ \Rightarrow 2(\mu+1)y = \lambda(k+\mu+1), 2x_{k+2+\mu} &= \lambda, \ldots, 2x_n = \lambda \end{aligned}$$

We solve first for λ and then for y and the x_i's.

$$\lambda = \frac{4}{\frac{(k+1+\mu)^2}{\mu+1} + n - k - 1 - \mu}$$

$$y = \frac{2^{\frac{k+\mu+1}{\mu+1}}}{\frac{(k+1+\mu)^2}{\mu+1} + n - k - 1 - \mu}$$
$$x_{k+2-\mu} = \cdots = x_n = \frac{2}{\frac{(k+1+\mu)^2}{\mu+1} + n - k - 1 - \mu}$$

Finally, the cost of this point is

$$f(y^*, x^*_{k+2-\mu}, \ldots, x^*_n) = \frac{4}{\frac{(k+1+\mu)^2}{\mu+1} + n - k - 1 - \mu}$$

It can be easily shown that this quantity is maximum when $\mu = n - k - 1$, that is when all the constraints are tight. In this case the cost is $4(n-k)/n^2$. Observe that this cost is bigger than the cost of Case 1. To finish the optimization we also need to check the value of f on the boundary of its domain. The boundary is where some of the x_i's are zero. This is equivalent of changing n in the above problem. Therefore we just need to optimize the cost with respect to n. Define $h(n) = 4(n-k)/n^2$ and differentiate

$$h'(n) = \frac{4n^2 - 4(n-k)2n}{n^4} = 0 \Rightarrow n = 2k$$
$$h''(2k) = -\frac{1}{2k^3} < 0$$

Therefore the maximum value of h is $h(2k) = 1/k$. This means that picking $k = 1/(\varepsilon^2\delta)$ brings the error (the function defined on the beginning of this proof) down to δ. ■

With all the above proven, we are ready to state the algorithm and combine the lemmas to get the desired result.

Algorithm 1 Edge Estimator

Input: $G_1, G_2, \varepsilon, \delta$

1: $k \leftarrow \Theta(\varepsilon^{-2}\delta^{-1})$
2: Construct Q consisting of the k highest degree vertices of G_1.
3: Query for the neighborhood of each $v \in Q$.
4: **return** $\hat{m} \leftarrow \sum_{(u,v)\in E(G)} \mathbf{1}(u \in Q(G_1) \vee v \in Q(G_1)) + \frac{1}{1-p^2} \sum_{e \in G[V\setminus Q]} \mathbf{1}(e \in E(G_2))$

Theorem 4.1. *Under the assumption that G has at least $\Theta(\varepsilon^{-2}\delta^{-1})$ vertices of degree higher than $24 \log n/(1-p)$, the estimator $\hat{m}$ described in Algorithm 1 which uses two samples and $k = \Theta(\varepsilon^{-2}\delta^{-1})$ queries satisfies $\mathbb{P}_{G_1,G_2\sim\mathcal{P}_G}(|\hat{m}-m| > \varepsilon m) \leq \delta$.*

Proof. Use Chebysev's inequality and each one of the lemmas

$$\begin{aligned}\mathbb{P}_{G_1,G_2\sim\mathcal{P}_G}(|\hat{m} - m| > \varepsilon m) &\leq \frac{\mathrm{Var}(\hat{m})}{\varepsilon^2 m^2} \\ &\leq \frac{c(p)\,\mathbb{E}_{G_1\sim\mathcal{P}_G}\left[\sum_{u\in V\setminus Q(G_1)} d^2_G(u)\right]}{\varepsilon^2 m^2} \qquad \text{(Lemma 4.1)}\end{aligned}$$

$$\leq \frac{a(p)c(p)\sum_{u\in V\setminus Q^*} d_G^2(u)}{\varepsilon^2 m^2} \qquad \text{(Lemma 4.1)}$$

$$\leq \delta \qquad \text{(Lemma 4.3)}$$

Where for the last step we set $\delta' = \delta/a(p)c(p)$ and applied Lemma 4.3 with δ' instead of δ. ■

We conclude with some remarks. First, we note that queries allowed to overcome the lower bound of $\Omega(\log(\varepsilon^{-2}\delta^{-1}))$ we had when only samples were used, as the algorithm now needs only two samples. The second is that we avoided explicitly writing the dependence on the parameter p in the expressions throughout our analysis. It is interesting to rewrite the query complexity of the estimator in terms of all the parameters to highlight the trade off between the number of samples and the number of queries. Suppose that we have N samples available and we use their union to reduce p to p^N as we did before. The trade-off is that each sample reduces exponentially the number of queries required.

Remark 4.1. *Algorithm 1 can be converted so that is takes N samples and $\Theta\left(\frac{p^N}{\varepsilon^2\delta}\right)$ queries.*

A note on the usage of two samples is that it makes the analysis very simple. If we instead used only the first sample G_1 in (4.3) in both terms, then the resulting estimator would not be unbiased anymore. To see that, let $Q(G_1)$ be the query set. Conditioned on the event $Q(G_1) = q$, the probabilities for each vertex to be hidden are not necessarily p anymore (for example if G is regular the query set will contain mostly visible vertices, and thus the probability that another vertex is hidden will be more than p).

4.3 Estimation of Triangles

In this section we extend the previous results to include triangle estimation. The analogues of Claims 4.1, 4.2 hold, with the former now having $T = \omega(n^2)$ as a requirement in order for the probability of error to be $o(1)$. However this is not very usefull because it is a quite strong requirement. For example, power law graphs do not meet this requirement as it was shown in [GvdHSS18] that the number of triangles in these graphs with parameter $\gamma \in (2, 3)$ is $\Theta(n^{\frac{3}{2}(3-\gamma)})$.

Now the optimal query strategy would be to ask for the true neighborhood of vertices that have many incident triangles. The only major difference is that now we cannot determine reliably from the sample graph which are these critical vertices because "triangle degrees" $\lambda(v)$ are not well preserved in the sample, as we may have graphs like the one of Figure 4.1 where, if edge (u, v) becomes hidden, it is imposible to find out which is the best vertex to query.

Definition 4.1. *The triplet of vertices (u, v, w) is called a wedge, if $(u, v), (v, w) \in E$.*

However, wedges are well preserved in the samples in the same way as degrees do, because for every pair $(u, v) \in V^2$ we have that $|\{w \in V \mid (u, w), (w, v) \in E\}|$

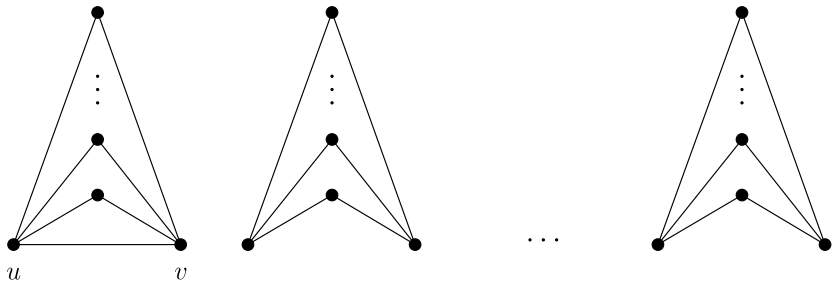

Figure 4.1. A worst case underlying graph.

either remains the same in the sample graph or becomes a binomial random variable. For that reason we will use wedges to determine which vertices to query, ending up in an estimator with a guarantee of the form $\mathbb{P}(|\hat{T} - T| > \varepsilon W) < \delta$, where W is the total number of wedges in the underlying graph. That guarantee is worse than the one we were initially aiming for (which had the error expressed in terms of T and not W) but we believe that the difficulty is inherent to the model. This guarantee is not firstly introduced here as it has been used in other works [SPK13].

Notation: We denote by T the total number of triangles and by W the total number of wedges. Let $\lambda(v)$ be the number of triangles adjacent to vertex v and $\lambda(u, v)$ the number of triangles that contain edge (u, v). Also let $w(v)$ be the number of wedges centered to vertex v, that is $w(v) = \binom{d(v)}{2}$, and $w(u, v)$ be the number of wedges that begin on u and end on v ((u, v) does not have to be an edge). We will also use subscripts when it is not clear what graph is under examination.

Suppose that a set Q of vertices is queried and the true neighborhood for each $v \in Q$ becomes known. We will assume that this set Q is deterministicaly decided and does not depend on the sample, while later we will drop this assumption. Divide the triangles of the underlying graph into three sets: $T_0 = \{t \in T(G) \mid t \text{ has no vertices in } Q\}, T_1 = \{t \in T(G) \mid t \text{ has one vertex in } Q\}$ and $T_{\geq 2} = \{t \in T(G) \mid t \text{ has 2 or 3 vertices in } Q\}$. The estimator's value for the sample G_s is defined as

$$\hat{T} = |T_{\geq 2}| + \frac{\sum_{t \in T_1} \mathbf{1}(t \in T(G_s))}{1 - p^2} + \frac{\sum_{t \in T_0} \mathbf{1}(t \in T(G_s))}{1 - 3p^2 + 2p^3}$$

It is easy to see that the above estimator is unbiased. We calculate its variance. The first term is not random and thus it does not contribute to the variance. Let $\hat{T}_e$ be the second term, which regards counting the edges that close triangles with one vertex from Q each time, and $\hat{T}_t$ be the final term which counts the triangles that have no vertex in Q. $\mathrm{Var}[\hat{T}] = \mathrm{Var}[\hat{T}_e] + \mathrm{Var}[\hat{T}_t] + \mathrm{Cov}(\hat{T}_e, \hat{T}_t)$. Denote by $\binom{Q}{2}$ the set of undirected pairs of vertices inside Q.

Lemma 4.4. *The variance of the estimator $\hat{T}$ is $\mathrm{Var}[\hat{T}] = \mathrm{Var}[\hat{T}_e] + \mathrm{Var}[\hat{T}_t] + Cov(\hat{T}_e, \hat{T}_t)$ where each term is bounded as follows (c_1, c_2, c_3 depend only on p and are specified in the proof):*

$$\mathrm{Var}[\hat{T}_t] \leq c_1(p) \sum_{u \in V \setminus Q} w_G^2(u) + c_2(p) \sum_{(u,v) \in \binom{V}{2} \setminus \binom{Q}{2}} w_G^2(u, v)$$

$$\mathrm{Var}[\hat{T}_e] \leq c_3(p) \sum_{u \in V \setminus Q} w_G^2(u)$$
$$\mathrm{Cov}(\hat{T}_e, \hat{T}_t) \leq \sqrt{\mathrm{Var}[\hat{T}_t]\,\mathrm{Var}[\hat{T}_e]}$$

Proof. Denote by G' the graph $G \setminus Q$, that is, the graph G after all vertices from Q are removed. Using the formula for the variance we derived in Section 3.6.1 we have that

$$\begin{aligned}
\underset{G_s \sim G}{\mathrm{Var}}(\hat{T}_t) &= \underset{G_s \sim G}{\mathrm{Var}}\left(\frac{|T(G_s \setminus Q)|}{1 - 3p^2 + 2p^3}\right) \\
&\leq \frac{6p^3(1-p)^3 \sum_{u \in V \setminus Q} \lambda_{G'}^2(u) + p^2(1-p)^2(1-2p)^2 \sum_{(u,v) \in E(G')} \lambda_{G'}^2(u,v)}{(1 - 3p^2 + 2p^3)^2} \\
&\leq \frac{6p^3(1-p)^3 \sum_{u \in V \setminus Q} \lambda_{G}^2(u) + p^2(1-p)^2(1-2p)^2 \sum_{(u,v) \in E(G')} \lambda_{G}^2(u,v)}{(1 - 3p^2 + 2p^3)^2} \\
&\leq c_1(p) \sum_{u \in V \setminus Q} w_G^2(u) + c_2(p) \sum_{(u,v) \in \binom{V}{2} - \binom{Q}{2}} w_G^2(u,v)
\end{aligned}$$

We now show that the variance of the term $\hat{T}_e$ is not bigger than the variance of $\hat{T}_t$. The term $\hat{T}_e$ is an edge counting term and thus its variance has the form of sum of squared degrees regarding the graph $G'' = (V \setminus Q, E'')$ where $E'' = \{(u,v) \in E(G) \mid \exists w \in Q \ : \ (u,v,w) \in T(G), u \notin Q, v \notin Q\}$.

$$\mathrm{Var}[\hat{T}_e] \leq c_3(p) \sum_{v \in V \setminus Q} d_{G''}^2(v) \leq c_3(p) \sum_{u \in V \setminus Q} \lambda_G^2(u) \leq c_3(p) \sum_{u \in V \setminus Q} w_G^2(u)$$

The covariance term can be bounded using Cauchy-Scrwarz inequality. ■

The optimal query strategy. Here we describe how the query set is defined. The query set contains the k highest degree vertices. Also, it contains both endpoints of the k pairs of vertices $(u,v) \in V^2$ that have the biggest $w_{G_1}(u,v)$.

Lemma 4.5. *Let Q^* be the optimal query set. It suffices for its size k to be $\Theta(\varepsilon^{-2}\delta^{-1})$ in order to have*

$$\frac{\sum_{u \in V \setminus Q^*} w_G^2(u)}{\varepsilon^2 W^2} \leq \delta \quad \textit{and} \quad \frac{\sum_{(u,v) \in \binom{V}{2} - \binom{Q^*}{2}} w_G^2(u,v)}{\varepsilon^2 W^2} \leq \delta$$

Proof (sketch). Exactly the same with that of Lemma 4.3. The similarity is due to the fact that $\sum_{u \in V} w_G(u) = W$ and $\sum_{(u,v)} w_G(u,v) = W$. Therefore, instead of degrees we have wedge counts and instead of edges we have total number of wedges. ■

Now we do the analysis for the setting where we are given two independent samples $G_1, G_2 \sim \mathcal{P}_G$ and we use the first sample to decide the query set $Q(G_1)$ according to the strategy described and the second sample to calculate the value of the estimator.

Lemma 4.6. *Let Q^* be the optimal query set of size k and $Q(G_1)$ the query set of size k calculated from the first sample. Under Assumption 1 stated for wedges instead of degrees, there exists a constant $a = a(p)$, depending only on p, such that*

$$\mathop{\mathbb{E}}_{G_1 \sim \mathcal{P}_G} \left[\sum_{u \in V \setminus Q(G_1)} w_G^2(u) \right] \leq a \sum_{u \in V \setminus Q^*} w_G^2(u)$$

$$\mathop{\mathbb{E}}_{G_1 \sim \mathcal{P}_G} \left[\sum_{(u,v) \in \binom{V}{2} \setminus \binom{Q(G_1)}{2}} w_G^2(u,v) \right] \leq a \sum_{(u,v) \in \binom{V}{2} \setminus \binom{Q^*}{2}} w_G^2(u,v)$$

Proof (sketch). Similar to Lemma 4.2. For the quantity $\sum_{u \in V \setminus Q} w_G^2(u)$ note that $w_G(u) = \binom{d_G(u)}{2}$ and thus only the degrees are involved just like in edge estimation. For the other sum $\sum_{(u,v) \in \binom{V}{2} - \binom{Q^*}{2}} w_G^2(u,v)$ note that if at least one of the vertices of a pair (u,v) is hidden, then the quantity $w_{G_1}^2(u,v)$ is binomialy distributed with population $w_G^2(u,v)$. This is the same behavior that degrees have, which was the only thing we exploited in the proof of the previous section. ■

Theorem 4.2. *For the estimator $\hat{T}$ with two samples G_1, G_2 and $k = \Theta(\varepsilon^{-2}\delta^{-1})$ queries as described above it holds that*

$$\mathop{\mathbb{P}}_{G_1, G_2 \sim \mathcal{P}_G} (|\hat{T} - T| > \varepsilon W) \leq \delta$$

Proof (sketch). Use Chebysev's inequality and then Lemmas 4.4, 4.5, 4.6 and also Cauchy-Scrwarz inequality. ■

Chapter 5

Conclusion

5.1 Remarks

We have introduced our noise model and presented efficient estimators for edge counting and triangle counting. Already from our examination of the case of triangles, where the core of our analysis is very similar to that of edges, it becomes clear that our approach is somewhat unified and thus its generalization for the case of other subgraph induced graphs, such as C_4 or cliques seems promising.

5.2 Future Directions

Apart from the extension of our results to other induced subgraphs, it would also be pleasing to resolve the issue of whether just one sample instead of two suffices for our estimation task. Intuitively, the claim that one sample is enough seems reasonable as the amount of information contained about the underlying graph remains roughly the same. Our analysis was based on two samples to avoid technical difficulties, but it would be interesting to investigate whether there exist more serious obstacles when only one sample is available for both the selection of the query set and the calculation of the output value.

Another direction would be to study other properties different than subgraph counting which are global, such as the diameter or the average distance. Instead of deriving an estimator for these properties, first they could be examined in a property testing setting, where one needs to find an algorithm to recognize if the graph has diameter or average distance which exceeds a certain threshold, say $\operatorname{poly}\log n$.

An interesting premise that is required as subroutine in some counting algorithms is sampling of edges uniformly at random, that is, the task of outputing an edge of the graph which is selected uniformly at random. This has been recently explored in the setting where standard type queries are the only way of accessing the graph [ER18, ERR19]. It would be interesting to find similar algorithms for our model.

Bibliography

[AB02] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[AKK18] Sepehr Assadi, Michael Kapralov, and Sanjeev Khanna. A simple sublinear-time algorithm for counting arbitrary subgraphs via edge sampling. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[AS66] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

[AYZ97] Noga Alon, Raphael Yuster, and Uri Zwick. Finding and counting given length cycles. *Algorithmica*, 17(3):209–223, 1997.

[B+16] Albert-László Barabási et al. *Network science*. Cambridge university press, 2016.

[BBCG08] Luca Becchetti, Paolo Boldi, Carlos Castillo, and Aristides Gionis. Efficient semi-streaming algorithms for local triangle counting in massive graphs. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–24. ACM, 2008.

[BLM13] S Boucheron, G Lugosi, and P Massart. A non asymptotic theory of independence, 2013.

[CEK+15] Flavio Chierichetti, Alessandro Epasto, Ravi Kumar, Silvio Lattanzi, and Vahab Mirrokni. Efficient algorithms for public-private social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 139–148. ACM, 2015.

[Cra46] Harald Cramir. Mathematical methods of statistics. *Princeton U. Press, Princeton*, page 500, 1946.

[Csi67] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

[CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[CW90] Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. *Journal of symbolic computation*, 9(3):251–280, 1990.

[DM13] Sergei N Dorogovtsev and José FF Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. OUP Oxford, 2013.

[Duc16] John Duchi. Lecture notes for statistics 311/electrical engineering 377. *URL: https://stanford. edu/class/stats311/Lectures/full_ notes. pdf. Last visited on*, 2:23, 2016.

[ELRS15] Talya Eden, Amit Levi, Dana Ron, and C. Seshadhri. Approximately counting triangles in sublinear time. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, FOCS '15, pages 614–633, Washington, DC, USA, 2015. IEEE Computer Society.

[EM02] Jean-Pierre Eckmann and Elisha Moses. Curvature of co-links uncovers hidden thematic layers in the world wide web. *Proceedings of the national academy of sciences*, 99(9):5825–5829, 2002.

[ER18] Talya Eden and Will Rosenbaum. On sampling edges almost uniformly. In *1st Symposium on Simplicity in Algorithms (SOSA 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[ERR19] Talya Eden, Dana Ron, and Will Rosenbaum. The Arboricity Captures the Complexity of Sampling Edges. In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi, editors, *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 52:1–52:14, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[ERS18] Talya Eden, Dana Ron, and C Seshadhri. On approximating the number of k-cliques in sublinear time. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 722–734. ACM, 2018.

[F+47] Will Feller et al. Harald cramer, mathematical methods of statistics. *The Annals of Mathematical Statistics*, 18(1):136–139, 1947.

[Gol17] Oded Goldreich. *Introduction to property testing*. Cambridge University Press, 2017.

[Gra11] Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

[GRS11] Mira Gonen, Dana Ron, and Yuval Shavitt. Counting stars and other small subgraphs in sublinear-time. *SIAM Journal on Discrete Mathematics*, 25(3):1365–1411, 2011.

[GvdHSS18] Pu Gao, Remco van der Hofstad, Angus Southwell, and Clara Stegehuis. Counting triangles in power-law uniform random graphs. *arXiv preprint arXiv:1812.04289*, 2018.

[HL13] Pili Hu and Wing Cheong Lau. A survey and taxonomy of graph sampling. *arXiv preprint arXiv:1308.5865*, 2013.

[Hub17] Mark Huber. A bernoulli mean estimate with known relative error distribution. *Random Structures & Algorithms*, 50(2):173–182, 2017.

[LC12] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[LF06] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2006.

[LKJ06] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Physical review E*, 73(1):016102, 2006.

[LR06] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[LSBER18] Timothy LaRock, Timothy Sakharov, Sahely Bhadra, and Tina Eliassi-Rad. Reducing network incompleteness through online learning: A feasibility study. In *The 14th International Workshop on Mining and Learning with Graphs. http://www. mlgworkshop. org/2018/papers/MLG2018_ paper_ 40. pdf*, 2018.

[Mou10] Nima Mousavi. How tight is chernoff bound, 2010.

[MU17] Michael Mitzenmacher and Eli Upfal. *Probability and computing: randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[New03] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[NWS02] Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002.

[Rao92] C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics*, pages 235–247. Springer, 1992.

[SERGP17] Sucheta Soundarajan, Tina Eliassi-Rad, Brian Gallagher, and Ali Pinar. ε-wgx: Adaptive edge probing for enhancing incomplete networks. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 161–170. ACM, 2017.

[Ses15] C Seshadhri. A simpler sublinear algorithm for approximating the triangle count. *arXiv preprint arXiv:1505.01927*, 2015.

[SPK13] Comandur Seshadhri, Ali Pinar, and Tamara G Kolda. Fast triangle counting through wedge sampling. In *Proceedings of the SIAM Conference on Data Mining*, volume 4, page 5, 2013.

[TKM09] Charalampos E Tsourakakis, Mihail N Kolountzakis, and Gary L Miller. Approximate triangle counting. *arXiv preprint arXiv:0904.3761*, 2009.

[TKM11] Charalampos E Tsourakakis, Mihail N Kolountzakis, and Gary L Miller. Triangle sparsifiers. *J. Graph Algorithms Appl.*, 15(6):703–726, 2011.

[TKMF09] Charalampos E Tsourakakis, U Kang, Gary L Miller, and Christos Faloutsos. Doulion: counting triangles in massive graphs with a coin. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 837–846. ACM, 2009.

[Wal39] Abraham Wald. Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, 10(4):299–326, 1939.

[WF+94] Stanley Wasserman, Katherine Faust, et al. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[WS98] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[Yu97] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.