Stockholms universitet

Sociala nätverk och skärningsgrafer

Sun Rui Sun

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2010:11
Matematisk statistik
September 2010

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2010:11**,
http://www.math.su.se/matstat

Sociala nätverk och skärningsgrafer

Sun Rui Sun*

September 2010

Sammanfattning

Sociala nätverk kan definieras genom förhållanden mellan människor som har något gemensamt. Det finns många olika sociala förhållanden, och därmed många olika konstruktioner av sociala nätverk. Det är svårt att skaffa sig en exakt bild av hur individer skapar sociala nätverk. Dock kan man mäta vissa egenskaper i nätverken och sedan försöka formulera modeller som fångar dessa egenskaper. Syftet med denna uppsats är att beskriva en modell för sociala nätverk baserat på så kallade slumpmässiga skärningsgrafer. Först beskrivs hur en graf kan användas allmänt för att modellera ett socialt nätverk och sedan vilka empiriska egenskaper sociala nätverk har. Efter en kort beskrivning av den enklaste slumpgrafmodellen, Erdös-Renyi grafen och dess olämplighet att modellera verkliga nätverk med, definieras skärningsgrafer samt beskrivningar och härledningar av deras egenskaper.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: susu5222@student.su.se. Handledare: Maria Deijfen.

Sammanfattning

Sociala nätverk kan definieras genom förhållanden mellan människor som har något gemensamt. Det finns många olika sociala förhållanden, och därmed många olika konstruktioner av sociala nätverk. Det är svårt att skaffa sig en exakt bild av hur individer skapar sociala nätverk. Dock kan man mäta vissa egenskaper i nätverken och sedan försöka formulera modeller som fångar dessa egenskaper.

Syftet med denna uppsats är att beskriva en modell för sociala nätverk baserat på den så kallade *slumpmässiga skärningsgrafer.* Först beskrivs hur en graf kan användas allmänt för att modellera ett socialt nätverk och sedan vilka empiriska egenskaper sociala nätverk har. Efter en kort beskrivning av den enklaste slumpgrafmodellen, Erdös-Renyi grafen och dess olämplighet att modellera verkliga nätverk, definieras skärningsgrafer samt beskrivningar och härledningar av deras egenskaper.

Abstract

Social networks are defined through relationships between people who have something in common. There are many different social relations, and thus many constructions of social networks. It is difficult to get an exact picture of how individuals create social networks. However, certain properties of networks can be measured and one can then try to formulate models that capture these properties.

The purpose of this paper is to describe a model for social network based on so-called *Random intersection graphs*. First describing in general how a graph can be used to model a social network and the empirical properties of social networks. After a concise description of the simplste random model of Erdös-Renyi graph and its weakness to model the real network, defining the random intersection graphs, derivation and description of their properties.

Förord

Denna uppsats är ett självständigt arbete om 15 hp vilket leder till en kandidatexamen i matematisk statistik vid Matematisk Institutionen, Stockholm Universitet.

Jag vill rikta ett stort tack till min handledare Maria Deijfen på matematiska institutionen, för hennes ovärderliga tålamod och vägledning under arbetets gång. Även tack till Elin Henriksson och Emmelie Skogman för att de hjälpte mig med språket.

Innehåll

1. Inledning

1.1 Bakgrund

Komplexa nätverk är ett omfattande begrepp vilka har stor betydelse i dagens samhälle. Dessa väcker stort intresse inom olika forskningsområden. Komplexa nätverk består av såväl kommunikationsnätverk som naturvetenskapliga nätverk. Uppsatsen börjar med en beskrivning av de nätverk vi befinner oss i dagligen.

Sociala nätverk består av grupper av människor som har någon typ av kontakt med varandra eller som det finns ett samband mellan. Det kan vara vänskap mellan individer, samarbete mellan företag eller relationer mellan familjemedlemmar.

Informationsnätverk kan även kallas för "kunskapsnätverk". Ett sådant kan uppstå när flera personer citerar akademiska artiklar knutna till varandra. Ett annat exempel av betydelse är den så kallade World Wide Web (WWW), vilket är ett nätverk för webbsidor som innehåller informationer som kopplas ihop via hyperlänkar.

Teknologiska nätverk är kartlagda nätverk för att distribuera vissa varor eller resurser, sådana som information eller elektricitet, t.ex telenätverk eller leveransnätverk. Ett typiskt och välkänt exempel är Internet, en fysisk koppling mellan datorerna.

Vissa biologiska system kan betraktas som nätverk såsom metaboliska nätverk vilka visar proteins interaktion, neuron nätverk som visar hur nervceller bildas och utvecklas, och regulatoriska gennätverk som beskriver hur DNA- segment påverkar gener.

Trots att nätverk kommer från många olika områden, har de gemensamma egenskaper såsom tungsvansad gradfördelning och hög klustring(se avsnitt 2.3 för definitioner av dessa begrepp).

Fokusen för denna uppsats ligger på sociala nätverk.

1.2 Problemformulering och översikt

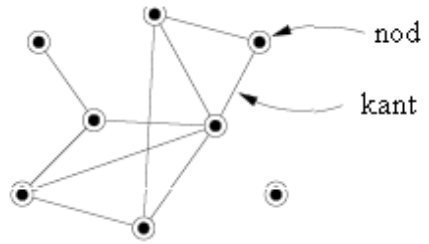
Den enklaste slumpgrafmodellen, ER-grafen har brister som inte kan undvikas när det gäller både gradfördelning och klustring. I denna uppsats beskrivs den så kallade slumpmässiga skärningsgrafen.

I kapital 2 beskrivs grundläggande terminologi för grafer samt empiriska egenskaper hos sociala nätverk. I kapital 3 definieras ER-grafer och deras egenskaper beskrivs. Kapital 4 ägnas åt slumpmässiga skärningsgrafer. Denna modell matchar de viktiga egenskaperna som kommer från empiriska studier.

2. Sociala nätverk och slumpgrafer

2.1 Grafteori

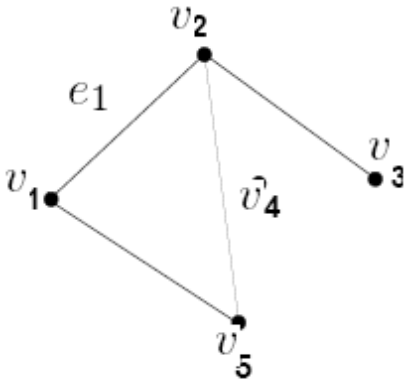
En graf består av små enheter “noder” och förbindelser mellan noder, så kallade “kanter”.



Figur 1. Ett enkelt nätverk med åtta noder och tio kanter

Noder och kanter är de grundläggande enheterna i en graf. Grafen kan nu användas för att beskriva empiriska nätverk t.ex sociala nätverk. Grafer kan delas upp i två kategorier: riktade grafer vilka har riktade kanter (t.ex telefonsamtal eller E-post mellan individer), och oriktade grafer med oriktade kanter.

Antag att det finns N noder som betecknas **V**= (V_1, …, V_N), och n kanter: **E**= (e_1, …, e_n), där e_k= (V_i ,V_j) vilken är koppling mellan noderna V_i och V_j , för k= 1,…., n, och i, j =1,…, N.



Figur .2. N=5, n=4

E= {e_1, e_2, e_3, e_4}= {e_1={V_1, V_2} , e_2={V_2, V_3}, e_3={V_1, V_5}, e_4={V_2, V_5}}

2.1.1 Grundläggande begrepp

För att undvika förvirring och få en klar bild av begreppen listas nedan de termer som kommer att användas i uppsatsen.

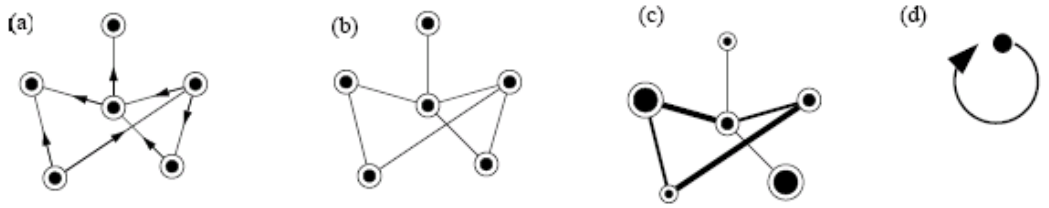
Noder
Den grundläggande enheten i nätverk.
Kant
Länkar mellan noder som kopplar ihop dem parvis. Kanter i samma graf kan ha olika längder, styrkor eller vikter.

Enkel graf

Graferna kan delas upp i två kategorier: riktade och oriktade. En enkel graf är en oriktad graf utan själv-loopar (dvs. utan kanter som börjar och slutar i samma nod) och med högst en kant mellan varje nodpar.

Bipartit graf

En bipartit graf består av två olika typer av noder. Kanterna sammankopplar endast noder av olika typer.



Figur. 3. (a) en riktad graf vilken bara har en riktning; (b) en oriktad graf med en typ av noder och en typ av kanter; (c) en bipartit graf av olika typer av noder och olika vikter av kanter; (d) en själv-loop.

Grad

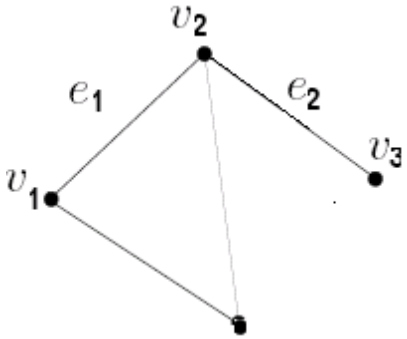
En nods grad ges av antalet kanter som är kopplade till denna nod. En riktad graf har både en inåt-grad och en utåt-grad (antalet inkommande resp. utgående kanter hos en nod) för varje nod.

Väg

En koppling från en nod till en annan nod i grafen. Det består av noder och kanter mellan dem.

Komponent

Två noder tillhör samma komponent om det finns en väg mellan dem. Noderna kan nå varandra i varje komponent med hjälp av en väg längs grafens kanter. En nod sägs vara *isolerad* om det utgörs av en egen komponent och en graf kallas *sammanhängande* om den bara har en komponent.



Figur.4. Vägen från nod V_1 till nod V_3 innehåller kant e_1, nod V_2 och kant e_2.

Avstånd

Avståndet mellan två givna noder är längden av den kortaste vägen mellan dessa noder, *medelnodavstånde*t i en graf är genomsnittet av avstånden i grafen.

Klustringskoefficienten

Klustringskoefficienten i en graf mäter i vilken utsträckning två noder som båda är kopplade till en given annan nod också är kopplade till varandra. Det finns två olika typer av

klustringskoefficient: global och lokal. Dessa beskrivs närmare i nästa avsnitt.

För att modellera nätverk används ofta slumpgrafer, dvs. grafer där antingen noderna eller kanterna uppstår slumpmässigt. Strukturer hos grafen(t.ex gradfördelningen och klustringen) blir då också slumpmässig. Vi ska undersöka detta för två olika slumpgrafmodeller i kapital 3 och 4.

2.2 Sociala nätverk

Ett socialt nätverk är en social struktur som består av individer av olika kön, olika nationaliteter, åldrar, inkomster etc, eller av organisationer. Dessa objekt representeras av noder och kanter vilka framställer någon form som samband mellan individer (bekantskap, vänskap eller något annat förhållande). Om det finns en kant mellan två noder V_i och V_j kallas de motsvarande individerna, *i* och *j*, *grannar*.

Exempel på sociala nätverk är vetenskapliga samarbetsnätverk, skådespelarnätverk och företagsstyrelser. Noderna i dessa nätverk är forskare, skådespelare resp. direktörer och kanter mellan dem består av forskare som skriver vetenskapliga artiklar ihop, skådespelare som uppträder i samma film respektive VD:ar som sitter i samma styrelse[4]. Ett annat flitigt använt nätverk är E-mailnätverket där noderna representeras av E-mailadresser och riktade kanter är mejl som skickas från en adress till en annan.

2.3 Egenskaper hos sociala nätverk

Många empiriska nätverk har gemensamma egenskaper. I detta arbete fokuseras på egenskaperna hos sociala nätverk. Det har visat sig att sociala nätverk (och även andra typer av nätverk) ofta har en potenslag i gradfördelning(en tung högersvans) och en hög klustring. Dessutom finns det en till intressant egenskap: medelnodavstånd (“Liten värld” fenomenet).

Även om det finns flera egenskaper som gradkorrelation, robusthet, gruppstruktur etc, berör uppsatsen egentligen bara gradfördelning och klustring. Andra egenskaper som inte förekommer i detta papper kan man hitta i [5].

2.3.1 Klustring

I ett socialt nätverk är sannolikheten att två av dina kompisar känner varandra betydligt större än sannolikheten att två slumpmässigt valda individer känner varandra. Detta mäts av klustringskoefficienten.

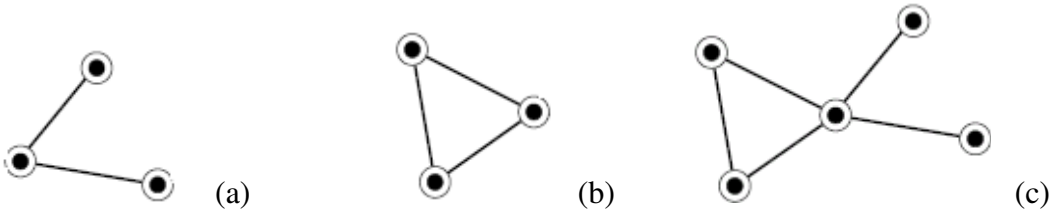
Av klustringskoefficienten har man två olika versioner: lokal och global. Den globala koefficienten ger en generell indikation av klustring över hela nätverket medan den lokala anger klustring för enskilda noder.

Vi betecknar den globala klustringskoefficienten C. Den är baserad på trianglar i nätverket. Om nod *u* kopplar till nod *v* och *v* är kopplad till nod *w* då bygger de tre noderna upp en triangel om även *u* kopplar till *w*. Om det inte finns en kant mellan *u* och *w*, bildar de en tripplett.

Klustringskoefficienten bestäms som:

$$C = \frac{3 \times \textit{antal trianglar i nätverket}}{\textit{antal tripletter i nätverket}} \tag{1}$$

Faktorn i täljaren multipliceras med tre för varje triangel har tre tripletter. C varierar mellan 0 och 1.



Figur.5. (a) en triplett; (b) en triangel har tre tripletter; (c) ett enkelt nätverk med en triangel och åtta tripletter, där den globala klustringskoefficienten **(def.(1))** är 3/8, de lokala koefficienterna **(def.(3))** är 1,1, 1/6, 0 och 0, och medelvärdet av samtliga lokala koefficienter **(def.(4))** är 13/30.

Det lokala värdet definieras genom klustringskoefficient C_i för en viss nod V_i :

$$C_i = \frac{\textit{antal trianglar förbundna med nod } V_i}{\textit{antal tripletter centrerade på nod } V_i} \tag{2}$$

För noder som bara har en granne eller är isolerade definieras $C_i = 0$.

Därefter finns även den genomsnittliga koefficienten. Den anger medelvärdet av de lokala klustringskoefficienterna:

$$\overline{C} = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{3}$$

Alla definitioner ovan är ur ett empiriskt perspektiv. I en stokastisk modell, dvs. en slumpgraf, kan man antingen använda väntevärden av uttrycken ovan eller se på den betingade sannolikheten att två noder är sammanlänkande med en kant om både är grannar med en gemensam nod.

2.3.2 Gradfördelning

Som ovan omnämnt är en nods grad antalet kanter som kopplar till denna nod. Det finns två olika slags grader i riktade nätverk: inåt-grad och utåt-grad, och bara en typ av grad i oriktade nätverk. I denna uppsats ska vi bara se på oriktade grafer.

En viktig observation är att gradfördelningen i många empiriska nätverk är höger-skev, vilket innebär att fördelningen har en lång högersvans av värden som ligger långt över medelvärdet. Detta orsakas av att ett fåtal noder har en betydligt högre grad än resterande noder i nätverket. Den höger-skeva gradfördelningen tenderar att visa upp en potenslag-svans, dvs. andelen P_k av noderna som har grad k ($0 \leq k \leq N-1$) uppfyller:

$$P_k \sim k^{-\tau} \tag{4}$$

För någon exponent $\tau > 1$. Här betyder $P_k \sim k^{-\tau}$ att $P_k / k^{-\tau}$ konvergerar mot en positiv konstant då $k \to \infty$. I empiriska nätverk ligger τ ofta mellan 2 och 3. Denna potenslagfördelning

kan grafiskt uppvisas med hjälp av en log-log plot:

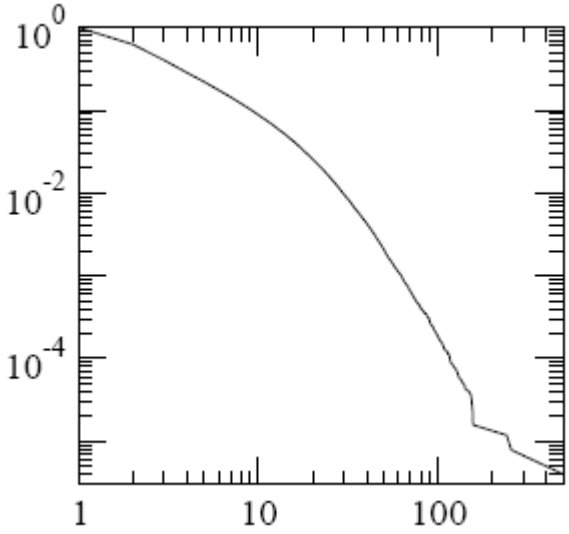
$$P_k \sim k^{-\tau} \Rightarrow \log P_k \sim -\tau \log k \tag{5}$$

Logaritmera båda sidor i (4), ser man att *log* P_k är en linjär funktion av *log k* med en negativ lutning som anges exponent -τ.

Alternativt kan graden även presenteras av den kumulativa fördelningsfunktionen som skrivs:

$$P_k = \sum_{k'=k}^{\infty} p_{k'} \tag{6}$$

Detta är sannolikheten att graden är större än eller lika med *k.* Fördelar med denna fördelning är att all osäkerheten i svansen minskar.



Figur.6. Kumulativ fördelning för ett samarbetes nätverk på loglog-skala. Den horisontal axeln är nodgraden, och den vertikala representerar den kumulativa sannolikhetens fördelning.

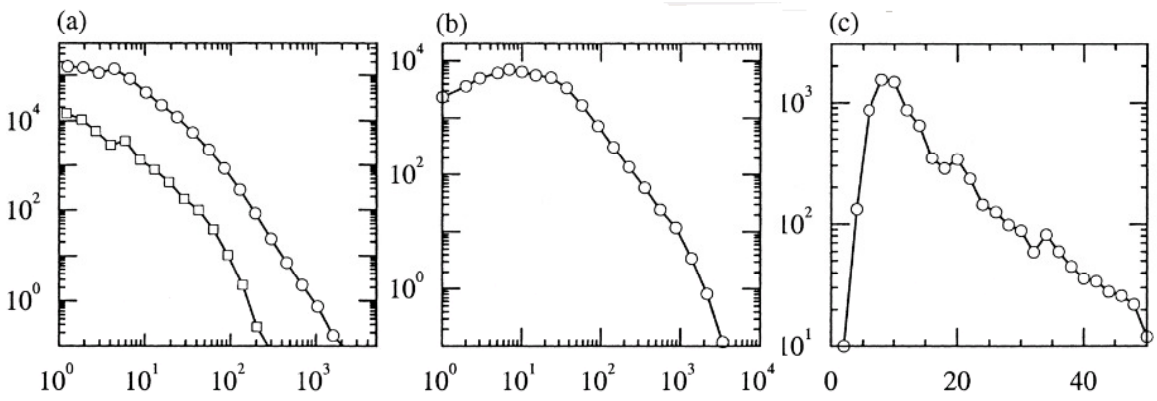
Potenslagen existerar även i den kumulativa fördelningen, istället för τ blir exponent τ-1:

$$P_k \sim \sum_{k'=k}^{\infty} k'^{-\tau} \sim k^{-(\tau-1)} \tag{7}$$

Somliga nätverk har en gradfördelning med exponentiell svans(ekvationen (8))[5], dvs.

$$P_k \sim e^{-\alpha k} \quad \text{för } \alpha>1 \tag{8}$$

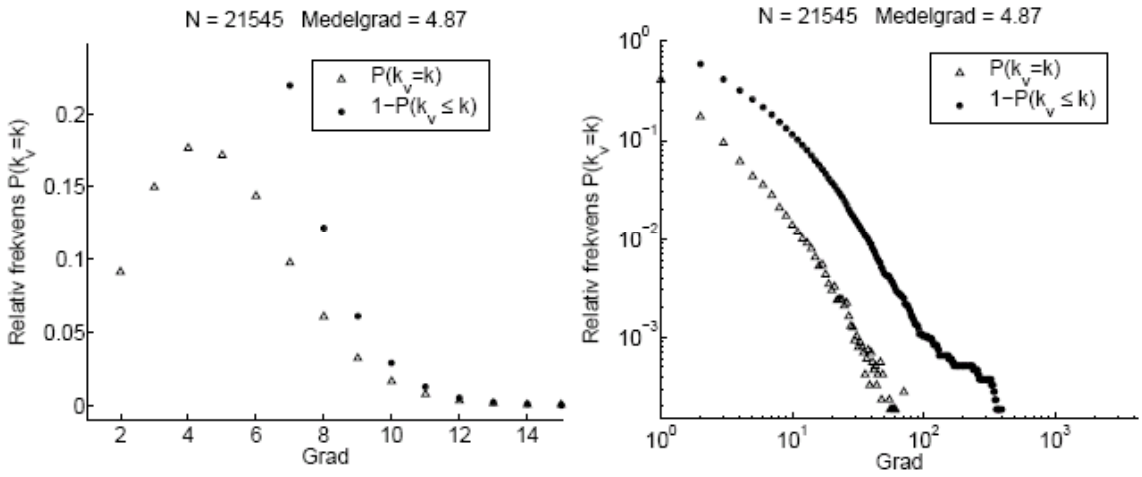
Logaritmera man (8) får man $\log P_k \approx -\alpha k$. En exponentiell svans innebär alltså att *log* P_k är en linjär funktion av *k*.



Figur.7. Tre olika typer av sociala nätverk: (a)Vetenskapligt samarbete; (b)Skådespelarnätverk; (c)Företagsstyrelse. Observera att (c) har en linjär horisontell axel medan (a) och (b) är logaritmiska på båda axlar.

Figur 7 visar det att (a) och (b) är mycket skeva och att (c) är mindre skev. En förklaring till skillnad mellan (a), (b) och (c) är att det är ett kontinuerligt arbete som företagsledare, medan samarbete mellan forskare och filmspelare bara har en engångskostnad, tar kortare tid och engångsarbete att skriva en artikel eller spelar in i en film[4].

Det är stor skillnad mellan ett nätverk med en exponentiellt avtagande gradfördelning och ett nätverk med en potenslagfördelning för graderna när det gäller strukturen. Majoriteten av noder i nätverk med exponentiella svansar har ungefär lika många kanter till sig, medan ett fåtal har lite fler eller lite färre. Men i potenslagnätverk har minoriteten av noder en stor mängd kanter medan de allra flesta noderna har få kanter.



Figur.8. Gradfördelning för ett Poisson-nätverk(till vänster) och ett potenslagnätverk(till höger) med lika många grader och genomsnittlig grad. Grafen till höger är i log-log skala och visar sig potenslag meden den till vänster visar en exponentiell svans[1].

2.3.3 "Liten värld" (" Small world")

Efter det mesta kända " världen är liten - experimentet", vilket beskriver hur nära folk är sammankopplade till varandra genom sociala kontakter konstaterade Stanley Milgram(1967)

att vi lever i en “Liten värld”.

Milgram gav olika personer brev och bad dem skicka till en given människa. Breven skickades från en deltagare till en annan deltagare tills de nådde fram till de tilltänkta personerna. Alla deltagare var inom samma bekantkretsen eller hade vissa samband. Ungefär en fjärdedel av breven kom fram till de angivna människorna, och i genomsnitt behövde breven bara levereras genom sex personer. Detta är ursprunget till hypotesen *“sex grader av åtskillnader (six degrees of separation)”* som säger att två godtyckliga personer i världen kan sammankopplas med högst fem personer som mellanled.

Efter Milgrams studie upptäckte man att begreppet “ Liten värld” nätverk existerar i flera verkliga system. Exempel på detta är filmskådespelarnätverket och författare till vetenskapliga artiklar.

2.4 Modeller

Dessa egenskaper för verkliga nätverk kommer från empiriska studier, vilket är normen för hur det kan skapa modeller för verkliga nätverk. Denna experimentella metod har genom intervjuer och mätningar lärt oss mycket om sociala strukturer. Men datasamlingen är begränsad i antal. Det behövs mycket arbete för att sammanställa uppgifterna. Det leder direkt till en önskan om att utveckla modeller som kan beskriva nätverken. En kvalificerad nätverksmodell ska inte vara för komplicerad. Dessutom ska den vara realistisk och ha likadana egenskaper som i verkliga nätverk (en tungsvansad gradfördelning och en hög klustring). Till exempel, den enklaste nätverksmodellen är Erdös-Renyi grafen, som beskrivs i nästa kapitel.

3. Erdös-Renyi grafen

I detta kapitel behandlas den enklaste slumpgrafmodellen, Erdös-Renyi modellen, vilken definierades av Paul Erdös och Alfred Renyi på 1950-talet. Enligt modellen sätts en kant mellan varje nodpar med samma sannolikhet, oberoende av andra kanter.

Till skillnad från föregående kapitel som byggde på empiriska nätverksstudier, ska ER-grafen användas till att modellera nätverk. Tyvärr visar det sig att vissa egenskaper i ER-grafen inte överensstämmer med empiriska resultat. Det är ändå intresserat att undersöka ER-modellen eftersom den har så enkel struktur och därmed är lätt att analysera.

3.1 Definition

Vanligtvis brukar ER- grafen betecknas *G (N, p)*, där *N* är antal totala noder i hela grafen och däribland sätts en kant mellan varje nodpar med samma sannolikhet *p,* oberoende av andra kanter. Antalet möjliga kanter *n* mellan *N* noder är:

$$n = \binom{N}{2} = \frac{N(N-1)}{2}$$

Antalet kanter i ER-grafen är alltså *Bin(n, p)* med väntevärde *np*. Parametern p varierar mellan 0 och 1. Grafen kommer att ha flera kanter om p ökar och färre kanter om p minskar.

3.2 Egenskaper i ER-grafen

3.2.1 Gradfördelning

En nods grad beror på mängder av kanter som ansluter till denna nod. Ju fler kanter en nod ansluter till desto högre grad blir det. I ER-graf är graden för en given nod binomialfördelad med parametrar *N-1* och *p, Bin(N-1, p).* Den förväntade graden är alltså *(N-1)p*.

Beteendet av ER-grafen studeras ofta då N går mot oändligheten, och $p=\lambda/(N-1)$ för någon parameter $\lambda>0$. Då får man en approximation till *Binomial-fördelning* från *Poisson-fördelning* med parameter λ.

Andelen noder med grad k ges alltså asymptotiskt av

$$P_k = \frac{e^{-\lambda} \cdot \lambda^k}{k!} \qquad (9)$$

och väntevärdet blir λ. Att skala kantsannolikheten med *N-1* gör alltså att den förväntade graden hålls konstant oberoende av befolkningsstorleken i det sociala nätverket.

De verkliga sociala nätverken brukar, som ovan nämnts, visa upp en potenslag i gradfördelningen. Men gradfördelning i ER-grafen som konvergerar mot Poisson fördelning varierar inte så mycket. Den har en exponentiell svans. Detta överensstämmer inte med de verkliga sociala nätverken, vilket tyder på att ER-grafen inte kan modellera verkliga nätverk.

3.2.2 Klustring i ER-grafen

Klustringen i Erdös-Renyi grafen är väldigt låg, vilken är ytterligare en anledning till att grafen inte kan användas till att modellera sociala nätverk med hög klustring. Den låga klustringen indikerar att det finns få trianglar i grafen. Genom se på den betingade sannolikheten att två noder är grannar, givet att de har en gemensam granne, kan man få en uppfattning om klustringen. Här är kanterna oberoende så att den betingade kantsannolikheten är lika med den obetingade när $N\to\infty$, $p=\lambda/(N-1)\approx\lambda/N$ som går mot 0 då $N\to\infty$, finns det alltså väldigt få trianglar i grafen.

ER-grafen innehåller alltså för få trianglar, och kan därför inte beskriva sociala nätverk som har hög klustring. Grafen misslyckas med att fånga de tungsvansade gradfördelningarna som man har sett i empiriska nätverk.

Således kan man att säga att den klassiska ER- grafen inte kan betraktas som en bra modell för sociala nätverk. Detta leder direkt till att man önskar ett bättre nätverksmodell som har liknande egenskaper som de verkliga nätverken. I följande kapitel presenteras en mer realistisk nätverkmodell: slumpmässiga skärningsgrafer.

4. Slumpmässig skärningsgraf

I detta kapitel beskrivs en annan slumpgrafmodell den s.k. slumpmässiga skärningsgrafen. Den har sin grund från en bipartit graf. Ett typiskt exempel på situationer där denna modell är lämplig är så kallad *tillhörighetsnätverk.* Samarbetsnätverket av forskare och filmskådespelare och nätverket av styrelse VD:ar är alla exempel på tillhörighetsnätverk.

Ett tillhörighetsnätverk består av två typer av noder, den ena är enskilda individer som skådespelare, författare och VD:ar, och den andra representerar olika grupper som filmer, artiklar eller styrelser. Kanterna går endast mellan noderna och grupperna, vilket betyder att t.ex författare bara sammankopplas med artiklar och tvärtom att artiklar enbart ansluter till författare. Grafen visar de artiklar som en viss författare har skrivit och samtidigt vilka författare som har deltagit i en artikel. En ny graf kan sedan definieras för enbart författare, där två författare är länkade om de har skrivit minst en artikel ihop.

4.1 Ursprunglig slumpmässig skärningsgraf

4.1.1 Definition

Den slumpmässiga skärningsgrafen, *G(N, M, p),* konstrueras med hjälp av en uppsättning *V= (V1, V2,…VN)* med *N* noder och en annan uppsättning med *M* element. En bipartit graf *B(N, M , p)* genererar först genom att varje given nod sammankopplas med varje givet element oberoende med sannolikhet *p*.

Grafen, *G (N, M, p)* som härrör från *B(N, M, p)* definieras sedan på *V* genom att två noder V_i och V_j sammankopplas om och endast om det finns ett element *a* så att både V_i och V_j är grannar till *a* i *B(N, M, p).*

Som en modell för sociala nätverk, där noderna i *V* kan betraktas som enskilda individer och elementen som sociala grupper, kan *G(N, M, p)* tolkas som att två individer är länkade om de delar minst en grupp.

4.1.2 Struktur

För att få en intressant struktur, väljas att $M=[N^{\alpha}]$ för $\alpha>0$. Låter D_i beteckna graden för nod V_i. Sannolikheten att två individer inte är i samma grupp är $(1-p^2)^M$. Således blir kantsannolikhet för noden V_i: $1-(1-p^2)^M$. Med hjälp av Taylorutveckling får vi att $(1-p^2)^M=(1-Mp^2+0(M^2p^4))$. Den förväntade graden blir alltså:

$$E[D_i]=(N-1)(1-(1-p^2)^M)=(N-1)(Mp^2+0(M^2p^4)) \tag{10}$$

För att begränsa den förväntade graden då $N\to\infty$, låter vi $p=\gamma N^{-(1+\alpha)/2}$ för någon konstant $\gamma>0$. Då fås:

$$E[D_i]=(N-1)N^{\alpha}(\gamma N^{-(1+\alpha)/2})^2 \to \gamma^2 \tag{11}$$

I en slumpmässig skärningsgraf spelar α en stor roll i gradfördelning med ovanstående *p*. Stark[6] visar hur gradfördelning beter sig när $\alpha<1$, $\alpha=1$ och $\alpha>1$:

(1). $\alpha<1$, då konvergerar gradfördelningen till en punkt massa vid 0;
(2). $\alpha=1$, då konvergerar gradfördelningen mot en summa av Poissonvariabler;
(3). $\alpha>1$, då konvergerar gradfördelningen till en Poissonfördelning.

Både (2) och (3) visar sig vara asymptotiskt Poisson. Detta tyder på att den ovanstående modellen inte kan användas för modellering av gradfördelning i sociala nätverk.

För att få fram en icke-Poisson gradfördelning, introduceras en generalisering av den ursprungliga modellen genom att oberoende associera en slumpmässig vikt till noderna. Vi ska också begränsa oss till fallet $\alpha=1$ i denna uppsats, eftersom det är detta som visar sig leda till en graf med realistisk gradfördelning och klustring.

4.2 Generalisering av modell

4.2.1 Definition

Denna modell, som kan betecknas *G (N, M, F)*, definieras på följande sätt[2]:

Tag en mängd $V= (V_1,\ldots,V_N)$ med *N* noder och en annan mängd med $M=[\beta N]$ element för $\beta>0$. Låt $\{W_i\}$ vara en i.i.d talföljd av positiva slumpvariabler med fördelning F med antagande att väntevärdet $E[W_i]$ är ändligt och lika med 1. Låt $\gamma>0$ och sätt:

$$p_i = \gamma W_i N^{-1} \tag{12}$$

Definiera en bipartit graf *B(N, M, F)* genom att, för varje nod V_i, dra en kant oberoende till varje element med sannolikhet p_i. *G(N, M, F)* konstrueras sedan genom att dra en kant mellan två noder i *V* om och endast om de har en gemensam granne i *B(N, M, F).*

I sociala nätverk kan viktparametern *W* exempelvis tolkas som ett mått på individers sociala aktiviteter. Noderna med stora vikter indikerar att man är med i många grupper och därigenom skaffar många sociala kontakter.

4.3 Gradfördelning

Antag att väntevärde för F är ändligt. Då kommer den asymptotiska förväntade graden hos noden V_i, betingat på W_i att vara lika med $\beta\gamma^2 W_i$:

Proposition 1. Låt D_i beteckna graden av en given nod $V_i \in V$ i en slumpmässig skärningsgraf *G(N, M, F)* med $M= [\beta N]$, och p_i definieras av (12). Om F har ändligt väntevärde(lika med 1), då har vi:

$$E[D_i \mid W_i] \rightarrow \beta\gamma^2 W_i \qquad \text{då } N\rightarrow \infty \tag{13}$$

Bevis: Låt $p_{\{ij\}}$ beteckna sannolikheten att det finns en kant mellan noderna V_i och V_j. Denna sannolikhet ges av:

$$p_{ij} = 1-(1-p_i p_j)^{\beta N} \approx 1-(1-\beta N p_i p_j) = \beta N p_i p_j \tag{14}$$

Väntevärdet av graden för V_i, betingade på vikten W_i, är summan över j≠*i* av den förväntade kantsann olikheten till nod V_j. Detta blir:

$$\begin{aligned} E\left[D_i \middle| W_i\right] &= \sum_{i\neq j} E\left[p_{ij}\right] = \sum_{i\neq j} E\left[\beta N p_i p_j\right] = \sum_{i\neq j} E\left[\beta N \gamma W_i N^{-1} \gamma W_j N^{-1}\right] \\ &= \sum_{i\neq j} \beta\gamma^2 W_i E\left[W_j\right] N^{-1} = \beta\gamma^2 W_i \sum_{i\neq j} E\left[W_j\right]\frac{1}{N} = \left\{E\left[W_j\right]=1\right\} = \beta\gamma^2 W_i \sum_{i\neq j}\frac{1}{N} \\ &= \beta\gamma^2 W_i \frac{N-1}{N} \to \beta\gamma^2 W_i \qquad \text{då } n\to\infty \end{aligned}$$

Påståendet är bevisat.

Vi ser nu på gradfördelningen:

Sats1. Låt D_i beteckna graden för en given nod V_i i en slumpmässig skärningsgraf *G(N, M, F)* med $M=[\beta N]$ och p_i som i (12). Fördelningen för D_i, betingat vikten W_i konvergerar mot fördelningen för ett $Po(\beta\gamma W_i)$ fördelat antal oberoende $Po(\gamma)$ variabler.

Bevis: Vi visar satsen för $i = 1$. Låt D_1 beteckna graden av noden *1* och *L* antalet grupper individ *1* är med i. Då är $L \sim Bin(\beta N p_1)$, vilket är approximativt $Po(\beta\gamma W_1)$(eftersom $p_1 = \gamma W_1 / N$) . Låt:

$$\mathbf{1}_j = \begin{cases} 1 & \text{om det finns minst en grupp som både V1 och Vj tillhör} \\ 0 & \text{annars} \end{cases}$$

Fördelningen för kantindikator $\mathbf{1}_j$ betingad på $\{W_j\}$ och *L* är approximativt $Be(Lp_j)$. För stora *N*, graden D_1 kan skrivas som $D_1 = \sum \mathbf{1}_j$ för j = 2,…n, och den sannolikhetsgenererande funktionen ψ_{D1} för D_1 blir:

$$\begin{aligned} \psi_{D_1}(t) &= E\left[t^{D_1}\right] = E\left[t^{\sum_{j=1} 1_j}\right] \\ &= E\left[E\left[t^{\sum_{j=1} 1_j} \middle| \{W_i\}, L\right]\right] = E\left[E\left[\prod_{j=2}^{N} \psi_{1_j} \middle| \{W_i\}, L\right]\right] \\ &= E\left[\prod_{j=2}^{N}\left(1-Lp_j+Lp_j t\right)\right] = E\left[\prod_{j=2}^{N}\left(1+(t-1)Lp_j\right)\right] \\ &= E\left[e^{\log\left(\prod_{j=2}^{N}\left(1+(t-1)Lp_j\right)\right)}\right] = E\left[e^{\sum_{j=2}^{N}\log\left(1+(t-1)Lp_j\right)}\right] \end{aligned}$$

Där, t ∈ [0, 1]. Enligt Taylorutveckling $\log(1+x)\approx x \Rightarrow \log(1+(t-1)Lp_j) \approx (t-1)Lp_j$,

$$\psi_{D_1}(t) = E\left[e^{(t-1)L\sum_j p_j}\right] = E\left[\left(e^{(t-1)\sum_j p_j}\right)^L\right]$$

$$= \psi_L\left(e^{(t-1)\sum_j p_j}\right) = E\left[\left(1+p_1\left(e^{(t-1)\sum_j p_j}-1\right)\right)^{\beta N}\right]$$

$$\to e^{\beta\gamma W_1(e^{(t-1)\gamma}-1)} \qquad då \quad n\to\infty$$

Låt $Z_1, ..., Z_L$ vara en följd oberoende $Po(\gamma)$ variabler och $L \sim Po(\beta\gamma W_1)$ oberoende av följden. Sätt $Z \overset{d}{=} Z_1 + Z_2 + , ..., + Z_L$, den genererande funktionen för Z är:

$$\psi_{D_1}(t) = E\left[\left(\psi_{Z_1}(t)\right)^L\right] = \psi_L\left(\psi_{Z_1}(t)\right) = e^{\beta\gamma W_1(e^{(t-1)\gamma}-1)}$$

Vi har visat att den genererande funktionen för D_1 konvergerar mot: $e^{\beta\gamma W_1(e^{(t-1)\gamma}-1)}$. Det följer från sats VI.5.1 och III.2.1 i Gut[3] att D_1 konvergerar i fördelning mot Z, och satsen är bevisad.

Genom att sätta vikten på noderna, kan man åstadkomma en potenslagfördelning för grader. Om viktfördelningen F följer en potenslag har de båda samma exponent, därigenom lyckas en tungsvansad gradfördelning genomföras som i det verkliga sociala nätverket. Därför är skärningsgrafen en kvalificerad modell till nätverksstudie när det gäller gradfördelning. Vi avslutar med att beskriva klustringen.

4.4 Klustring

Klustringen i skärningsgrafen är också betingat på vikten. Låt H_{ij} beteckna händelsen att V_i och V_j delar minst en grupp. Vidare definieras $C_{i,j,k}$ som kantsannolikheten mellan noderna V_i och V_j givet att båda två är anslutna till noden V_k, betingat på vikter: $C_{i,j,k} = P_n(H_{ij}/H_{jk}, H_{ik})$. Detta är den betingade sannolikheten att de tre noderna V_i, V_j och V_k bygger upp en triangel, betingat på att två av de tre kanterna redan existerar.

Den asymptotiska klustringen $C_{i,j,k}$, anges i följande sats:

Sats 2. Låt i, j, k vara tre distinkta noder i $G(N, M, F)$ med $M=[\beta N]$ och $p_i = \gamma W_i/N$, och antag att F har ändligt väntevärde, då fås :

$$C_{i,j,k} \xrightarrow{P} (1+\beta\gamma W_k)^{-1} \tag{15}$$

$C_{i,j,k}$ beror på antalet grupper som V_k medverkar i vilken i genomsnitt är $\beta\gamma W_k$. När $\beta\gamma W_k$ är stort minskar sannolikheten att V_i och V_j är länkade till V_k via samma grupper(och alltså även själv är länkade). Tvärtom ökar sannolikheten om $\beta\gamma W_k$ är litet. Den förväntade asymptotiska klustringen $E = \left[(1+\beta\gamma W_k)^{-1}\right]$ kan varieras mellan 0 och 1 genom att justera β och γ.

Slutsatsen som kan dras genom att välja F- viktfördelning med den önskade exponenten och också lämpliga parametrar β och γ , fås det slutligen en graf med ett givet värde av klustring och en potenslag gradfördelning med samma exponent som i viktfördelningen, vilken överensstämmer med den tungsvansade gradfördelningen från empiriska studier. På grund av

detta kan skärningsgrafen tillämpas då man modellerar sociala nätverk.

5. Slutsatser

I början av uppsatsen beskrivs olika nätverk och analyser av deras egenskaper från de empiriska undersökningarna. Dessa visar att verkliga sociala nätverk har en höger-skev(tungsvansad) gradfördelning och en hög klustringskoefficient, i motsats till Poisson gradfördelning och den låga klustringen hos Erdös-Renyi modellen. En annan grafmodell, den slumpmässiga skärningsgrafen, presenterades. I denna modell kan man få tungsvansade gradfördelningar och hög klustring, vilket stämmer överens med de empiriska observationerna.

Referenser

[1] Bovin ‚Johan. ***"komplexa nätverk". Smittskyddsinstitutets rapportserie 1:2003***
[2] Deijfen, M. & Kets, W., (2009): ***"Random intersection graphs with tunable degree distribution and clustring". Probability in the Engineering and Informatinal Sciences 23, 661-674***
[3] Gut, A., (1995): ***An intermediate course in probability. Springer. sid 169 & sid 62***
[4] Newman, M.E.J., Watts, D.J & Strongatz, S.H., (2002): ***"Random graph models of social networks". PNAS 99, 2566-2572***
[5] Newman, M.E.J., (2003): ***"The structure and funktion of complex networks". SLAM review 45, 167-256***
[6] Stark D., (2004): ***"The vertex degree distribution of Radom intersection graphs". Combinatorics Probability & Computing 24, 249-256***