Matematisk statistik
Stockholms universitet

Modeller för sociala nätverk

Grzegorz Czernik

Examensarbete 2010:1

Postadress:

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm
Sverige

Internet:

http://www.math.su.se/matstat

Matematisk statistik
Stockholms universitet
Examensarbete **2010:1**,
http://www.math.su.se/matstat

Modeller för sociala nätverk

Grzegorz Czernik*

April 2010

Sammanfattning

Här beskrivs från den första enklaste slumpgrafmodellen till mer avancerade och realistiska modeller för sociala nätverk som används idag. De viktigaste modellerna för sociala nätverk definieras och modellernas egenskaper lyfts fram. Först ägnas en stor del åt att beskriva de viktigaste egenskaperna som olika slags nätverk genom olika empiriska studier har visat sig ha: tungsvansad gradfördelning, hög klustring och litet medelnodavstånd. Hur bra en modell är går hand i hand med hur väl den lyckas avbilda dessa egenskaper. Senare vid genomgången av själva modellerna (Erdos-Rényi modellen, Smallworld-modeller, Bipartita modeller, Preferential attachment modeller) framhålls hur väl respektive egenskap uppfylls och därmed också indirekt hela modellens lämplighet. Det visar sig att Erdos-Rényi modellen endast har medelnodavståndet som en godtagbar egenskap. Small-world-modeller uppfyller också medelnodavståndet, och dessutom också klustringen. Bland bipartita modeller finns en stor variation. En bipartit modell deVnerad i kapitel 5.3 uppfyller både gradfördelningen och klustringen, men något resultat för medelnodavståndet finns inte tillgängligt. Preferential attachment modeller uppfyller medelnodavståndet och gradfördelningen.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post: grcz7733@student.su.se. Handledare: Maria Deijfen.

Abstract

This report provides an overview of social network modelling. The most common models of social netoworks are defined and the properties of the models are highlighted. The most important empirical properties of real networks are described: heavy-tailed degree distribution, large clustering and small average path length. It is then investigated how the proposed models succeed in capturing these properties.

The models that are described are the Erdős-Rényi graph, small-world models, models based on bipartite graphs and preferential attachment models. As for the Erdős-Rényi graph, it captures the small average path length but is not realistic when it comes to the degree distribution and the clustering. The small-world models capture the small average path length and the clustering, but not the degree distribution. As for the bipartite models, there is a large variation. A particualar bipartite model that is described turns out to capture both the degree distribution and the clustering, but no results for the average path length are yet available. Preferential attachment models gives rise to realistic degree distributions and has small average path length.

Förord

Detta arbete är ett examensarbete i matematisk statistik om 30 poäng och leder till en magisterexamen i matematisk statistik vid Stockholms universitets matematiska institution.

Jag vill tacka Maria Deijfen för hennes synpunkter, råd och hängivenhet under hela arbetets gång.

Innehåll

1 Introduktion

Syftet med denna rapport är att beskriva ett antal modeller för sociala nätverk samt kommentera deras lämplighet.

* * *

Studier av nätverk har vuxit fram efter att man inom olika ämnesområden intresserat sig för sociala nätverk. Ett socialt nätverk består av individer och deras relationer till varandra. Det är just relationer som definierar ett socialt nätverk.

Studier av *komplexa* nätverk startade i slutet av 1990-talet i och med att mer kraftfulla datorer gjorde det möjligt att studera stora verkliga nätverk empiriskt och undersöka deras egenskaper. Att nätverken är komplexa syftar på att de är stora samt att deras struktur är varken helt regelbunden eller helt oregelbunden. De flesta nätverk som studeras idag är komplexa och brukar delas in i huvudgrupperna sociala nätverk, informationsnätverk samt teknologiska och biologiska nätverk.

Vad som framkommit av studier är att många olika nätverk har likartade egenskaper, trots att de inte tillhör samma grupp, utan kan till och med komma från mycket olika ämnesområden. Dessa likheter har ökat intresset för nätverk då forskningen kan finna en större tillämpbarhet. Att nätverken på så sätt skulle ha en gemensam gåta att lösa motiverar den stora mängd teoretiska forskning vars mål är att kunna formulera modeller för att generera nätverk där man har kontroll över dessa egenskaper. Egenskaper som har studerats särskilt mycket är till exempel gradfördelning, klustring och avstånd mellan noder (individer i sociala nätverk).

En grupp av nätverk som studerats mycket, och som är det huvudsakliga föremålet för denna rapport, är sociala nätverk. Dessa nätverk är intressanta av många olika skäl, bland annat därför att sociala relationer påverkar beteendet hos processer som äger rum på nätverket och som i sin tur påverkar oss individer. Exempel på sådana processer kan vara spridning av smitta eller information. Vill man kunna studera och förutsäga beteendet hos sådana processer är det intressant att ha en modell som kan generera ett nätverk med egenskaper som påminner om egenskaperna hos ett verkligt socialt nätverk. Detta är en svår uppgift eftersom de nätverk man är intresserad av är typiskt sett mycket stora och omöjliga att skaffa sig en exakt bild av – de är komplexa.

För att modellera ett nätverk används naturligen en graf och den oregelbundna stukturen hos ett socialt nätverk antyder att en slumpgraf (det vill säga en graf vars konstruktion involverar någon typ av slumpmekanism) är att föredra framför en helt regelbunden graf. Nätverk har studerats långt innan man började inse nyttan av sannolikhetsteorin inom området. Det var först efter att Paul Erdős och Alfréd Rényi definierade slumpgrafen år 1959 som sannolikhetsteorin fick, och ännu till idag har, en självklar plats i studier av nätverksmodellering. Att en slumpmekanism används i modelleringen medför att man får resonera ur ett sannolikhetsteoretiskt perspektiv för att förutsäga modellens egenskaper.

Begränsningarna med nätverksmodellering är att modellerna inte får bli för komplicerade, att de måste vara realistiska samt att nätverken som man vill återskapa är svåra att få en exakt bild av.

Rapporten skulle kunna delas upp i två delar, där det först kommer en "studiedel" och sedan en "modelleringsdel". Den första delen skulle vara det nästföljande kapitel 2 som beskriver verkliga nätverk och deras egenskaper, och den andra delen de resterande kapitlen som beskriver modellerna, som man försöker definiera på ett sådant sätt att de som sagt ska kunna generera ett simulerat nätverk som uppvisar liknande egenskaper som de verkliga nätverken.

2 Nätverk och grafer

Olika typer av nätverk, däribland sociala nätverk, kan modelleras med grafer. Genom empiriska studier av verkliga nätverk har det visat sig att att många olika typer av nätverk inte skiljer sig åt alltför mycket, tvärtom har man sett att de olika nätverken faktiskt har flera gemensamma egenskaper. Detta har lett till att man börjat tro att det kan finnas ett generellt svar på vilka egenskaper ett nätverk har. I denna rapport om sociala nätverk är därför mycket av det som är sant om sociala nätverk också sant för andra typer av nätverk, och vice versa. Detta samband är anledningen till att också de tas upp. Viktigt att komma ihåg är att i hela detta kapitel beskrivs endast *verkliga* nätverk, det vill säga nätverk som tillkommit genom att man samlat in och analyserat information från nätverk som finns omkring oss, se kapitel 2.3 för exempel på sådana verkliga nätverk. De resultat som man erhållit ur dessa empiriska studier tar man fasta på när man sedan försöker sig på att skapa modeller för de verkliga nätverken; man måste ju veta vad man vill återskapa. Detta innebär att den bästa modellen är just den som bäst stämmer överens med de empiriska resultaten.

Grafen som ska modellera ett nätverk är själva grunden i nätverksteorin och har därför en stor betydelse. I kapitel 2.1 beskrivs vad som menas med en graf ur ett rent matematiskt perspektiv, eftersom grafen kan ses som ett matematiskt verktyg i studier av nätverk. I kapitel 2.2 beskrivs själva modelleringen, det vill säga hur man tänker sig att grafer ska kunna representera nätverk. Här tas kan man säga själva grundidén upp till hela nätverksmodelleringen. Kapitel 2.3 låter oss ta del av nätverksteorins mångfald då det ges exempel på olika typer av nätverk som har studerats. Kapitel 2.4 handlar om den samlade information som man genom många olika studier har byggt upp om nätverkens egenskaper och som man sedan i de övriga kommande kapitlen bygger modeller utifrån. Till sist i kapitel 2.5, då man redan vid det stadiet vet en hel del om nätverk, ges en liten djupdykning i två sociala nätverk som studerats mycket.

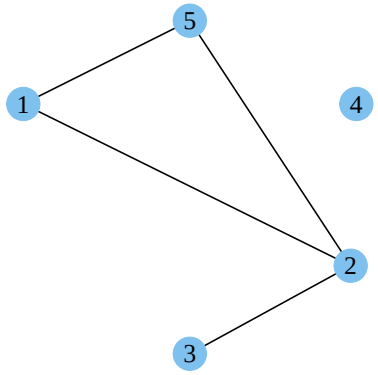
2.1 Grafteori

Grafer studeras inom *grafteorin*, en gren inom matematiken som utvecklades långt innan grafen fann sin tillämpning inom nuvarande nätverksteori. Det hela började med att Leonhard Eulers, en av alla tiders mest produktiva matematiker, visade år 1736 i en artikel att det klassiska matematiska problemet *Königsbergs sju broar* var olösligt. Detta anses vara det första resultatet inom grafteorin och sedan dess har området vuxit kraftigt.

En *graf* byggs upp av två olika slags beståndsdelar, *noder* och *kanter*. Noder illustreras som punkter och kanter som linjer som sammanbinder noderna, se figur 1 nedan. En nod i en graf kan vara allt från helt isolerad, det vill säga inte vara sammanbunden med någon annan nod, till sammanbunden med samtliga andra noder i grafen – totalt $N - 1$ stycken andra noder om grafen består av totalt N stycken noder. Om alla noder på detta sätt är sammanbundna med samtliga andra $N - 1$ noder, fås en graf där det går en kant mellan samtliga nodpar och grafen får det

maximala antalet kanter som den kan ha, vilket är $(N(N-1))/2$.

En graf skrivs formellt som G = {V, E}, där V är *nodmängden* och E är *kantmängden*. Således är V en mängd av N noder, V={ V_1, V_2,..., V_N}, och E är en mängd av kanter som visar hur kanterna sammanbinder noderna.



Figur 1. Exempel på en graf. En graf G = {V, E} består av de två mängderna V och E. Grafen ovan har $N = 5$ och kan skrivas som G={{1, 2, 3, 4, 5}, {{1, 2}, {1, 5}, {2, 3}, {2, 5}}}.

Man inser att en och samma graf kan representeras grafiskt på flera olika sätt och detta visar på att det som egentligen definierar en viss graf är endast antalet noder och hur de sammanbinds genom kanterna, och allt detta framgår i notationen. Men en grafisk representation av en graf fyller ändå sin funktion och är användbar om inte grafen är alltför stor. En stor graf är svårare att ta till sig och då försöker man kvantifiera istället, det vill säga beräkna nyckeltal från datat som bygger upp grafen.

Grafen ovan är endast det enklaste exemplet på en graf. Det finns dock många olika typer av grafer, samt också några speciella nyckeltal som man ofta räknar fram från grafer. Nedan följer en lista på många vanligt förekommande graftermer, se figur 2 för en illustrering av några av termerna.

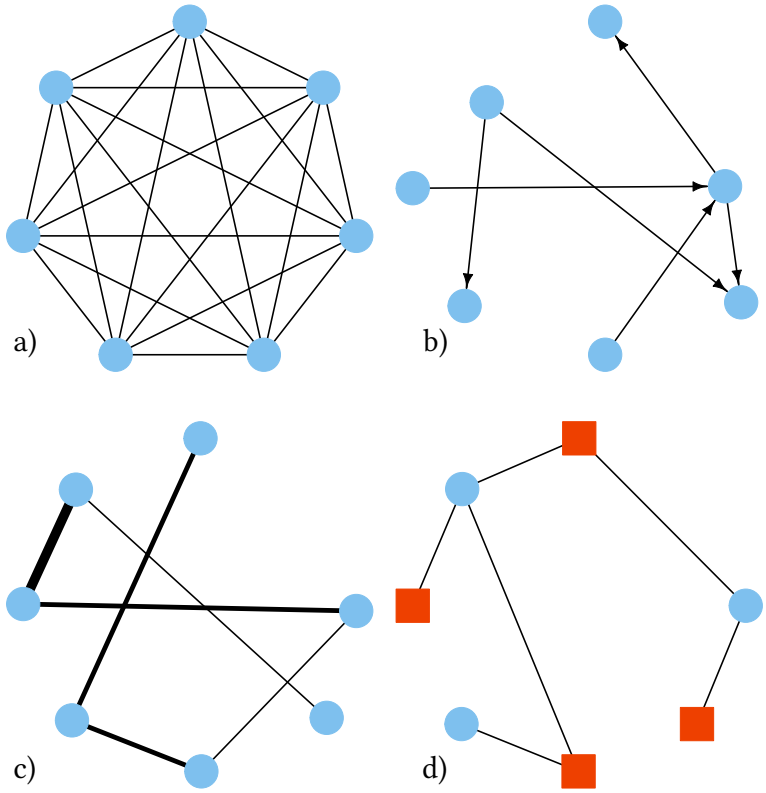
Termer inom grafteorin

- *Nod* : Punkt i grafen. Kallas också *hörn*.
- *Kant* : Linje mellan två punkter i grafen. Kallas också *båge*.
- *Granne* : Två noder är grannar om de sammanbinds med en kant.
- *Oriktad graf* : Kanterna har ingen speciell riktning utan kopplar helt enkelt ihop två noder.
- *Riktad graf* : Kanterna pekar i en viss riktning, vilket markeras med pilar.
- *Stig* : Ett sätt att gå mellan två noder genom att följa kanterna från nod till nod. Samma kant får inte användas två gånger. I en riktad graf får man bara gå i pilens riktning.

- *Cykel* : En stig som börjar och slutar i samma nod och som passerar andra noder högst en gång.
- *Avstånd mellan två noder* : Längden av den kortaste stigen genom grafen från en av noden till den andra. Det kan finnas och ofta också finns fler än en stig mellan två noder. Hur lång den kortaste stigen mellan två noder är definieras som det antalet kanter som finns längs den kortaste vägen som förenar noderna.
- *Grad* : Antal kanter som en nod har. För riktade grafer används in-grad och ut-grad för varje nod, vilka är antalet inkommande respektive utgående kanter.
- *Komponent* : Om det finns en stig mellan två noder tillhör noderna samma komponent. En komponent är således en mängd av noder där varje nod kan nås från varje annan nod.
- *Diameter* : Diametern av en graf är längden (räknat i antal kanter) av det längsta avståndet mellan två noder i grafen.
- *Densitet* : Anger tätheten av kanter i en graf. Densiteten beräknas som kvoten mellan antalet kanter i grafen och det största möjliga antalet kanter i grafen

$$densitet = \frac{\text{antal kanter i grafen}}{\frac{N(N-1)}{2}}$$

- *Viktad graf* : Om kanterna har vikter, det vill säga positiva reella tal tilldelade. En viktad graf illustreras till expemel genom att kanterna får en siffra eller olika tjocklekar.
- *Enkel graf* : En graf som är oriktad, oviktad och utan loopar (kanter som går från en nod tillbaka till sig själv) och utan multipla kanter (två eller fler kanter som sammanbinder ett nodpar).
- *Komplett graf / fullständig graf* : En oriktad graf där det finns en kant mellan varje par av noder.
- *Bipartit graf* : Graf med två olika typer av noder och där kanterna går endast mellan de olika nodtyperna.
- *Delgraf* : En delgraf till grafen G = {V, E} består av en delmängd E_1 av kanterna och en delmängd V_1 av noderna (som måste innehålla alla hörn som ingår i någon av kanterna i E_1).



Figur 2. Variationer på grafer. a) en *komplett graf* . b) en *riktad graf* . c) en *viktad graf* med tre olika vikter. d) en *bipartit graf* . Den högsta grad som graf c) har är två, vidare har den också endast en komponent.

2.2 Modellering av nätverk med grafer

I detta avsnitt beskrivs hur modelleringen av nätverk med grafer är tänkt. Detta kan ses som en allmän idé och som är till hjälp för att förstå modellerna som beskrivs i de kommande kapitlen.

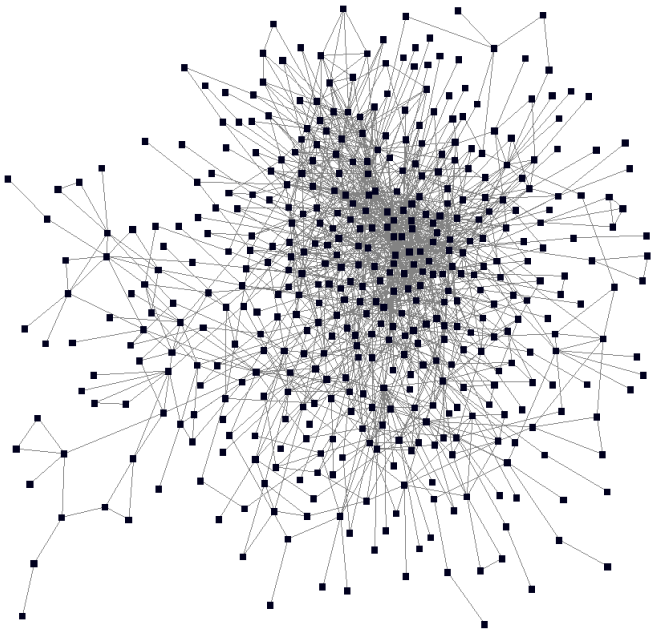
Grafer kan användas till att representera olika typer av nätverk som kan delas in i huvudgrupperna informationsnätverk, teknologiska nätverk, biologiska nätverk men också inte minst sociala nätverk – som ju är föremål för denna rapport. Studier av sociala nätverk handlar om individer och deras relationer med varandra. Exempel på sociala nätverk inkluderar bland annat vänskapsband mellan individer, företagsrelationer mellan företag och blandäktenskap mellan familjer (se kapitel 2.3 för fler exempel).

En sådan graf som illustreras i figur 1 skulle kunna beskriva ett mindre socialt nätverk genom att låta varje nod vara en individ och kanterna mellan dem skulle innebära intereaktioner eller kopplingar mellan individerna i nätverket i form av till exempel vänskapsband, släktskap, professionella bekantskaper eller geografisk närhet. Oftast används mycket större grafer eftersom nätverken som studeras ofta är stora. Se Tabell 1 under kolumnen N för exempel på storlekar av studerade nätverk.

Syftet med grafen är att försöka avbilda ett verkligt socialt nätverk och då kan man kanske finna inte bara en större, utan också en något mer anvancerad graf än den som visas ovan i figur 1 vara mer lämplig. Grafen i figur 1 är den enklast möjliga men har tyvärr endast begränsade möjligheter. En lite mer avancerad graf fås om man istället låter olika noder ha olika egenskaper, som den bipartita grafen i figur 2 d) med två olika nodtyper. Denna graf är lämplig för modellering av till exempel män och kvinnor som då representeras av de olika nodtyperna, då man tänker sig att män och kvinnor har olika egenskaper. En annan vanlig modifiering är att också låta olika kanter ha olika egenskaper. Ett exempel är en viktad graf (figur 2 c)) där kanternas vikter skulle kunna representera hur väl individerna känner varann. En tjockare kant, en kant med en större vikt, skulle kunna representera att dem två individer som kanten sammanbinder känner varandra bättre än två individer som sammanbinds med en tunnare kant. Vidare kan kanter vara riktade åt endast ett håll när kanten har en pil åt en viss riktning (figur 2 b)). Grafen är i så fall en riktad graf och används till exempel när man modellerar nätverk över telefonsamtal eller e-posttraffik.

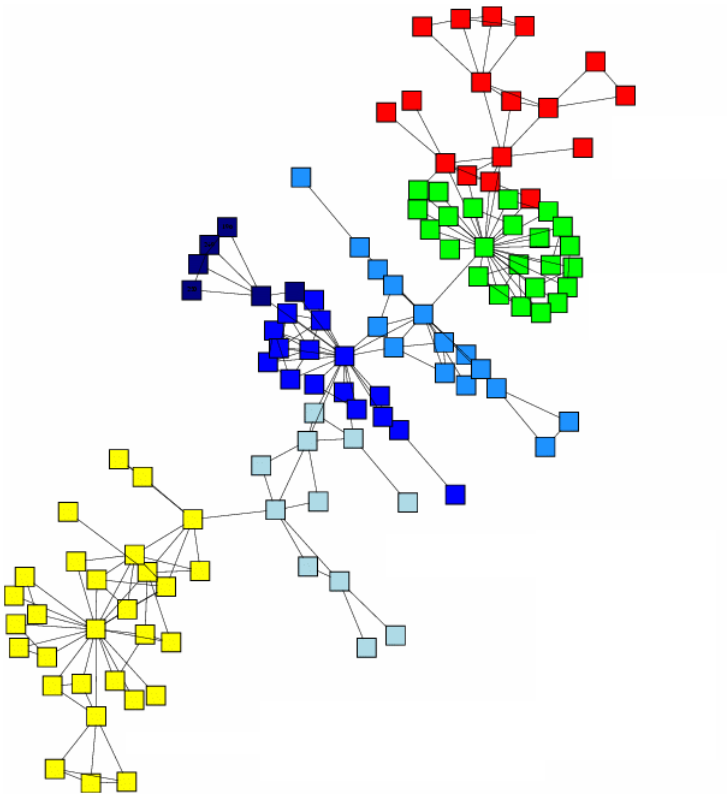
Nedan följer exempel som belyser lite mer i detalj några specifika nätverk och hur de representerats av grafer. Det främsta syftet är att bekanta sig med tankegångarna.

I figur 3 a) visas en illustrativ bild som helt enkelt visar hur nätverket över samarbeten mellan forskare skulle kunna se ut. Varje nod är en forskare och varje kant visar ett samarbete mellan två individer. Ett samarbete föreligger om forskarna samförfattat en artikel. Noder i mitten av grafen representerar individer som samförfattat artiklar med många forskare, medan noder längst ut representerar individer som ofta endast samförfattat en artikel. Ett sådant känt och mycket studerat samförfattarnätverk är samförfattandet av artiklar inom matematiken (kapitel 2.5.1 Erdőstal).



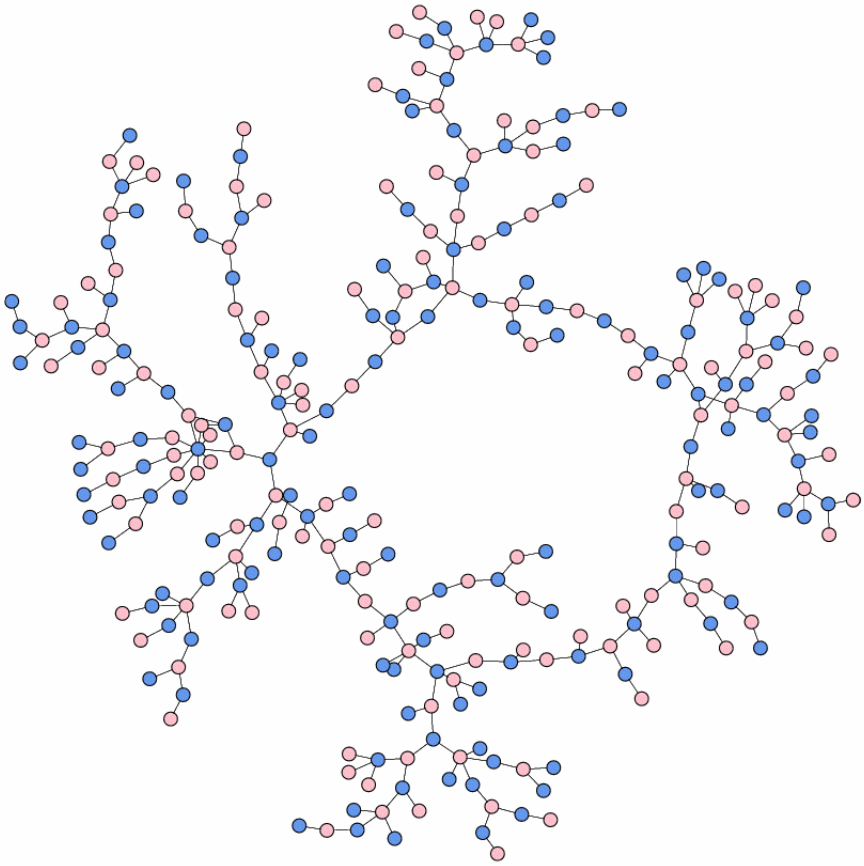
Figur 3. a) En illustrativ bild [8] på hur nätverket över samförfattandet av vetenskapliga artiklar skulle kunna se ut.

I figur 3 b) visas också samarbete mellan forskare. Även här representerar noder forskare och kanter visar ett samarbete mellan två individer. Men nu representeras forskarna med inte mindre än sju olika nodtyper, med olika egenskaper. Nodtyperna delar upp individerna i grafen efter inom vilket ämnesområde som individen i fråga är verksam. Samtliga nodtyper verkar ha en egenskap gemensam, och det är att de mycket oftare sammanbinds med noder av samma typ. Detta har sin naturliga förklaring i att forskare samarbetar i en mindre utsträckning över sitt ämnesområde än inom. Man väljer en grafs utseende beroende på vad man vill lyfta fram. Här förväntade man sig en sådan uppdelning och valde därför att representera grafen med olika nodtyper för att bättre kunna se denna uppdelning rent grafiskt. Dessutom buntade man ihop noder av samma typ. Om grafen endast skulle ha en nodtyp och om noderna inte skulle vara ihopbuntade skulle man säkerligen inte upptäcka denna starka uppdelning som man nu ser. En del studier av nätverk handlar om att med datorer försöka identifiera just sådana uppdelningar. Mellan nodtyperna kan förmågan att binda till andra noder variera, både till noder av samma typ och till andra typer. Till exempel binder endast en nodtyp till så många som tre olika nodtyper.



Figur 3. b) Graf [9] över samförfattarskap av artiklar mellan olika ämnesområden. De olika färgerna representerar skilda ämnesområden. Samförfattarskap över ämnesområderna är mindre förekommande än inom samma ämnesområde och gruppstrukturer (kapitel 2.4.4.4) framkommer tydligt. Grafen visar den största komponenten i Santa Fe Institutets samarbetsnätverk.

I figur 3 c) finns en bipartit graf som skulle kunna illustrera ett dejtingnätverk på ett gymnasium. Noder är individer och kanter visar dejtingrelationer mellan individpar. De två olika nodtyperna representerar kön och nodernas uppenbara egenskap är i detta fall att de endast kan sammanbindas mellan olika nodtyper. En annan egenskap skulle kunna vara skillnader i förmågan att sammanbinda till noder, det vill säga att individer av det ena könet överlag skulle ha haft fler dejtingrelationer på gymnasiet än det andra. Om grafen skulle vara viktad, kunde de olika vikterna representera hur länge dejtingrelationen varade.



Figur 3. c) En bipartit graf [9] som skulle kunna illustrera ett dejtingnätverk på ett gymnasium.

Som nämns i inledningen till detta kapitel används vid skapadet av grafmodeller för sociala, liksom andra typer av nätverk, resultat från empiriska studier där man tittat på vilka egenskaper som ett nätverk har i verkligheten. Dessa resultat används sedan vid skapandet av en modell som förhoppningsvis fångar upp dessa egenskaper. Vid modellering av nätverk används en *slumpgraf*. Anledningen till detta är att ett nätverk har en komplex struktur som beskrivs bättre av en graf vars konstruktion involverar någon typ av slumpmekanism, än av en deterministisk graf. Alla modeller i denna rapport är slumpgrafer. Vid skapandet av en modell (slumpmodell) handlar det om att definiera grundläggande lokala egenskaper i modellen, som till exempel sannolikheten att en kant dyker upp mellan två noder, som man sedan bygger hela

grafen efter. Man bygger alltså en stor graf genom att först definiera lokala egenskaper.

Hur ett visst nätverk ser ut i verkligheten beror på olika faktorer. Man är förstås intresserad av att få kännedom om dessa faktorer som påverkar nätverkets struktur för att finna stöd vid definierandet av de lokala egenskaper hos en graf. En lokal egenskap är i stort sett en mekanism som bestämmer sättet på vilket grafen byggs upp. En lokal egenskap tillskrivs ett visst fenomen i verkligheten.

Från början studerade man små regelbundna grafer och hade frågeställningar kring dem. Man kunde fråga sig vem som var den viktigaste personen i ett socialt nätverk, då en individ i ett litet nätverk kunde vara betydelsefull för hela nätverket. Nuförtiden studerar man stora grafer och frågeställningarna är mer av statistisk karaktär. Nu vill man till exempel ta reda på hur stor andel av kanterna måste tas bort för att dela upp grafen i minst två komponenter. Man kvantifierar datat med hjälp av sammanfattande nyckeltal (kapitel 2.4) då graferna blivit så stora att det blivit svårt att ta till sig dem på något annat sätt. Datorutvecklingen har självklart bidragit till att analysen av stort datamaterial och simuleringar av grafer blivit enklare.

Datainsamlingen från sociala nätverk är inte alltid så tillförlitlig, utan det är vanligt att insamlingen har problem med felaktigheter, att subjektivitet kan infinna sig och att urvalet inte är tillräckligt stort. Med undantag för ett fåtal fyndiga indirekta studier som Milgrams experiment (kapitel 2.4.1) sker datainsamlingen vanligtvis genom att fråga deltagarna direkt via frågeformulär eller genom intervjuer. Sådana metoder kräver mycket arbete och som en följd av detta blir storleken på nätverket som kan studeras begränsat. Datamaterialet som man använder sig av blir kanske inte heller objektivt besvarat, vilket beror på att det till exempel kan skilja sig mellan svaranden när det gäller definitionen av en vän.

Även om man gör stora ansträngningar för att eliminera felkällor av den typen är det allmänt accepterat att det finns stora och okontrollerade fel i de flesta av dessa sociala studier. Däremot kan man få mycket mer tillförlitlig information i de fall där olika typer av databaser finns tillgängliga. Exempel på nätverk som skapats ur en databas är nätverk över filmskådespelare (kapitel 2.5.2 Bacontal), där två skådespelare sammanbinds om dem medverkat i samma film. Information över skådespelare och film är noggrannt dokumenterat i filmdatabasen Internet Movie Database

`http://www.imdb.com/`

Liknande datamaterial kan hittas för att skapa nätverk över samförfattarskap bland forskare, där två forskare sammanbinds om dem har skrivit en eller flera artiklar tillsammans, samt nätverk över bolagsdirektörer, där två direktörer sammanbinds om dem sitter i samma styrelse.

2.3 Olika typer av verkliga nätverk

Det är som bekant möjligt att studera flera olika nätverk med nätverksteori. Nedan listas exempel på vanliga nätverk som man studerat samt till vilken grupp de tillhör. Vissa av nätverken är riktade eller bipartita. Datamaterialen för vissa av nätverken är också mer tillförlitliga än för andra, där nätverk byggda på komplett registerdata som till exempel nätverk över filmskådespelare, skådespelarnätverket, måste höra till de mest tillförlitliga. Kanske är detta anledningen till att de två mest studerade informationsnätverk är citationsnätverk och webben. De flesta nätverk växer också hela tiden, eller åtminstone så byts noderna och kanterna ut eller också så försvinner de. Detta är ytterligare en egenskap som noder och kanter kan ha i slumpgrafsmodeller.

Verkliga nätverk

- *Samförfattarskap* : Noder är forskare och kanter sammanbinder två forskare om dem har skrivit en artikel i samarbete. Nätverket är ett oriktat socialt nätverk.
- *Skådespelare* : Noder är filmskådespelare och kanter sammanbinder två filmskådespelare om dem har medverkat i samma film. Nätverket är ett oriktat socialt nätverk.
- *Telefonsamtal* : Noder är telefonnummer och avslutade telefonsamtal är kanter, riktade från den som ringer till den som svarar. Nätverket är ett riktat socialt nätverk.
- *Webben (World Wide Web)* : Noder är internetsidor och kanter är hyperlänkar. Vissa studier är gjorda på sajtnivå där internetsidor som tillhör samma sajt representeras av en och samma nod. Nätverket är riktat och följer hyperlänkarna, men i vissa fall har man använt nätverk med kanter som inte är riktade. Nätverket är ett informationsnätverk.
- *Citationsnätverk* : Noder är publicerade artiklar och riktade kanter är referenser till tidigare publicerade artiklar. Nätverket är ett riktat informationsnätverk.
- *Lingvistik* : Olika nätverk kan skapas. Noder kan vara ord och kanter sammanbinder två ord om dem finns brevid eller ett ord från varandra i en mening. Man kan också få ett nätverk över synonymer om noder är ord och kanter sammanbinder två ord om dem är synonymer. Nätverken är oriktade informationsnätverk.
- *Internet* : Man kan studera internet på två nivåer. Noder kan vara routrar och datorer och kanter är fysiska länkar mellan dem, eller så kan noder vara domäner, där varje domän innehåller hundratals routrar och datorer, och kanterna sammanbinder två domäner om det åtminstone finns en router som förbinder dem. Nätverket är ett teknologiskt nätverk.
- *Kraftnät* : Noder är generatorer, transformatorer och transformatorstationer och kanter är högspänningstransmissionsledningar. Nätverket är ett oriktat teknologiskt nätverk.

– *Ekologiska nätverk* : Noder är olika arter och kanter är predator-byte relationer mellan dem. Nätverket används av ekologerna för att kvantifiera interaktionerna mellan olika arter. Nätverket är ett riktat biologiskt nätverk.
– *Neurala nätverk* : Noder kan vara neuroner, och kanter kan vara synaptiska förbindelser. Nätverket är ett biologiskt nätverk.
– *Metaboliska reaktionsvägar* : Noder är substrar och produkter, och kanter sammanbinder ett visst substrat med en viss produkt om det finns en känd metabolisk reaktion som omvandlar substratet till produkten. Nätverket är ett oriktat biologiskt nätverk.
– *Proteinveckning* : Noder är olika tillstånd, konformationer, av ett protein, kanter sammanbinder konformationerna om dem kan erhållas från varandra genom en elementär förflyttning. Nätverket är ett biologiskt nätverk.

2.4 Egenskaper hos sociala nätverk

I detta avsnitt följer olika empiriska resultat från studier av verkliga nätverk. Dessa resultat är viktiga eftersom de ligger till grund när man börjar arbeta med att ta fram slumpgrafmodeller. Som redan påpekats har det observerats att många olika typer av nätverk tenderar att ha likartade egenskaper, vilket innebär att mycket av det som beskrivs gäller även för andra typer av nätverk. Nedan följer en kort förklaring av de viktigaste och mest studerade egenskaperna som man sett att verkliga nätverk har, innan en längre förklaring ges längre fram. Avsnittet avslutas med ytterligare några vanligt förekommande egenskaper.

Viktigaste egenskaper hos sociala nätverk

- *"Världen är liten"-fenomenet* : Det faktum att det genomsnittliga avståndet mellan noder tenderar att vara litet i förhållande till nätverkets storlek.
- *Klustring* : Att det finns många trianglar i nätverket. *Klustringskoefficienten* $C \in [0, 1]$ är ett mått på förekomsten av triangelstrukturer i nätverket. Sociala nätverk har en stor förekomst av trianglar, vilket ger en relativt hög klustringskoefficient.
- *Tungsvansad gradfördelning* : En grafs gradfördelning $\{p_k\}$ anger sannolikhetsfördelningen för nodgraderna. Verkliga nätverkens gradfördelning följer en potenslag $p_k \sim k^{-\gamma}$, där värdet på γ brukar vara $2 \leq \gamma \leq 3$.

2.4.1 Medelnodavstånd och "världen är liten"-fenomenet

Man har upptäckt vid studier av sociala nätverk, liksom vid andra typer av nätverk, att den kortaste vägen mellan de flesta nodpar i ett nätverk verkar bestå endast av ett fåtal kanter, trots att nätverket självt kan vara mycket stort, och det är detta som är känt som *"världen är liten"-fenomenet*, på engelska *"small-world effect"*.

Hela idén bakom "världen är liten"-fenomenet introducerades först av Stanley Milgram år 1967 efter att han utförde ett experiment vars syfte var att undersöka medelavståndet mellan två individer i en population. Experimentet genomfördes genom att Milgram valde individer i USA i städer som Omaha, Nebraska, Wichita och Kansas och gav dem till uppgift att skicka brev till en viss förbestämd målperson. Om de inte kände denna person direkt skulle de sända brevet vidare till en av sina bekanta som de bedömde skulle ha störst möjlighet att vidarebefordra brevet till målpersonen på snabbast möjliga sätt, det vill säga på ett sätt med så få mellanleder som möjligt. Ungefär en fjärdedel av breven nådde fram och de hade då passerat i genomsnitt sex personer. Experimentet har sedan gett upphov till begreppet *"sex grader av åtskillnad"*, på engelska *"six degrees of separation"*, som är en hypotes om att alla på jorden är sammanlänkade med varandra med som högst fem personer som mellanled.

Ett konkret exempel på ett nätverk som uppvisar ett "världen är liten"-fenomen ges i en studie [5] gjort över samförfattarskap av artiklar inom neurologi publicerade mellan 1991 och 1998. Nätverket har ett medelnodavståndet på $l = 6$. Ett annat exempel är webben. Nätverket är mycket stort och ett delnätverk på 325 729 noder har studerats [4] där det visade sig att medelnodavståndet var $l = 11.2$.

Numera har "världen är liten"-fenomenet studerats och verifierats direkt i ett stort antal olika nätverk. "Världen är liten"-fenomenet innebär bland annat att det också går fort att ta sig mellan noderna i många riktiga nätverk, eftersom det går fortare att hoppa ett fåtal steg, vilket blir fallet med fenomenet, än kanske 100 steg. Allmänt beräknas medelnodavståndet l mellan nodpar i ett nätverk som [10]

$$l = \frac{1}{\frac{1}{2}N(N+1)} \sum_{i \geq j} d_{ij} \tag{1}$$

där d_{ij} är avståndet mellan nod i och j, och där $d_{ii} = 0$. Definitionen av (1) är problematisk om nätverket har fler än en komponent, för i sådana fall finns det nodpar som inte har någon sammanbindande väg alls. Man brukar tilldela oändliga avstånd till sådana par, men då blir också värdet på l oändligt stort. För att undvika detta definieras vanligtvis l i sådana nätverk som medelnodavstånd mellan alla par som har en förbindelse. Par av noder där noderna inte tillhör samma komponent räknas alltså helt enkelt inte in i medelvärdet.

Matematiskt definieras "världen är liten"-fenomen ofta som att avstånden mellan noderna växer högst logaritmiskt med antalet noder N. Det är naturligt att tänka sig att avstånden mellan noderna är större i en stor population än i en liten, men som beskrivits ovan är ökningen i praktiken ganska långsam, och $\ln N$ är en funktion som ökar relativt långsamt med N. Måttet (1) på avstånd mellan noder är ett

empiriskt mått, det vill säga ett mått på observerat data. I en stokastisk modell är avstånden mellan noderna förstås stokastiska. Låt H_N beteckna avståndet mellan två noder som väljs slumpmässigt ur alla par av noder som är sammanlänkade med en stig. Modellen sägs ha "världen är liten"-egenskapen om

$$\lim_{N\to\infty} P(H_N \leq c \ln N) = 1 \tag{2}$$

för någon konstant c, det vill säga om N är stort så kan man med stor sannolikhet ta sig mellan två noder i högst $c \ln N$ steg.

2.4.2 Gradfördelning

Ett nätverk består oftast av noder med olika grader. Den lägsta grad en nod kan ha är ju noll och innebär att noden är helt isolerad. Den högsta grad en nod kan ha det är när den har en kant till samtliga andra noder i grafen, vilket är $N - 1$.

Man kan göra ett histogram över ett nätverk där man delat in noderna efter grader. Sedan kan man låta p_k beteckna *den andelen noder i nätverket som har graden* k. Detta definierar en sannolikhetsfördelning $\{p_k\}$ med $k = 0, 1, 2, ..., N - 1$, och där p_k är sannolikheten att en slumpmässigt vald nod har graden k.

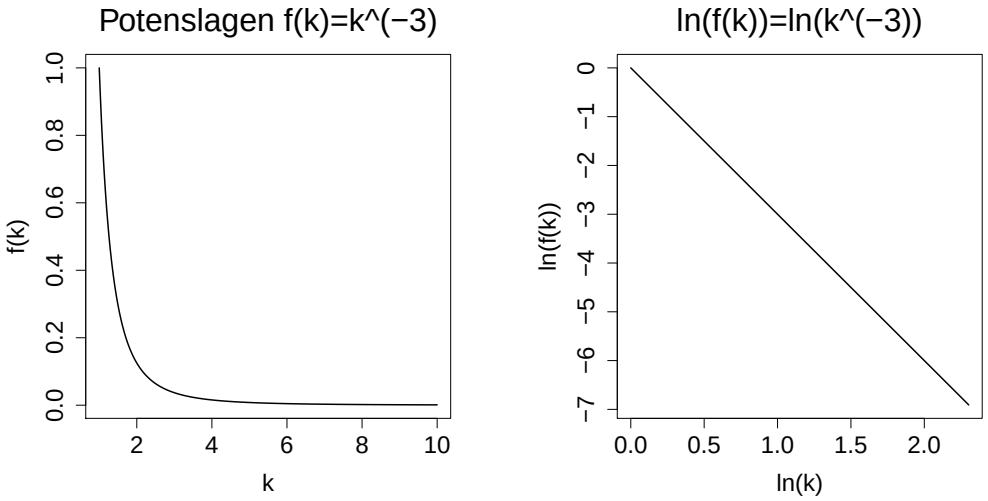
De flesta verkliga nätverk har en gradfördelning med en relativ lång högersvans med värden som är långt ifrån medelvärdet. En sådan skev gradfördelning fås om en stor majoritet av noderna har en låg grad men ett litet antal noder har en hög grad. En fördelning med en liknande lång högersvans som man funnit användbar när man försökt modellera gradfördelningen är *potenslagen*, som ges av

$$p_k \sim k^{-\gamma} \tag{3}$$

där γ är en konstant. Ekvationen innebär att $p_k/k^{-\gamma}$ konvergerar mot en positiv konstant då k går mot oändligheten. Ett typiskt värde på γ i modelleringar av verkliga nätverk är $2 \leq \gamma \leq 3$, se tabell 1.

Ett konkret exempel på ett nätverk vars gradfördelning följer en potenslag är ett nätverk över telefonsamtal. Nätverket är riktat och har två stycken gradfördelningar, en för in-grad och en för ut-grad. Långdistanta telefonsamtal ringda under en enda dag studerades [1] och det visade sig att gradfördelningen för in-grad respektive ut-grad följde en potenslag med exponenten $\gamma_{\text{in}} = \gamma_{\text{ut}} = 2.1$. Ett annat exempel på ett nätverk som följer en potensfördelning är återigen webben [4]. Delnätverket av webben som studerades gav att $\gamma_{\text{in}} = 2.1$ och $\gamma_{\text{ut}} = 2.45$.

Nätverk som följer en potenslag kallas ibland för *skalfria nätverk*.



Figur 4. *Till vänster* : exempel på potenslagen som verkliga nätverk följer. I detta fall är $\gamma = 3$. *Till höger* : samma potenslag i logaritmisk skala. Linjens lutning är -3.

Det absolut vanligaste sättet att grafiskt upptäcka en potenslagfördelning ur sitt data är med hjälp av en ln-ln plot (se högra bilden i figur 4). Det vill säga man plottar andelen noder p_k med grad k mot k, och logaritmerar båda axlarna. Ser man då en rät linje tyder det på att gradfördelningen följer en potenslag och linjens lutning anger exponenten.

Det är svårt att mäta en fördelning med sådan lång högersvans som de verkliga nätverken uppvisar. Men för att försöka göra det skulle man kunna tänka sig att göra som beskrivits ovan och göra ett histogram över nätverkets noder, men detta brukar resultera i att man får förlite data i svansen för att statistiken ska vara tillförlitlig. Men det finns två accepterade sätt att ta sig runt problemet. Det ena sättet är att konstruera ett histogram där de olika intervallängderna ökar exponentiellt med graden. Man skulle kunna ha ett histogram med intervallängder som delar in noderna efter graderna 1,2-3,4-7,8-15 och så vidare. Sedan delas det observerade datat i varje intervall med intervallets längd för att normalisera mätningarna. Detta är ett lämpligt sätta att konstruera ett histogram på ifall histogrammet ska plottas med en logaritmisk skala för grad, för att då förefaller bredden på intervallen vara lika breda. Intervallen blir bredare ju längre ut i svansen man går och detta medför att problemet med förlite data i svansen minskar. Det andra sättet att presentera grader på är genom att plotta den kumulativ fördelningsfunktion

$$P_k = \sum_{k'=k}^{\infty} p'_k \tag{4}$$

som är sannolikheten att graden är större eller lika med k. När man gör ett histogram genom att dela in värden i olika intervall tar man samtidigt bort alla skillnader mellan datapunkter som hamnar i samma klass. Den kumulativ fördelningsfunktion har inte detta problem utan all ursprunglig data kan representeras. Detta är en fördel med en sådan plot, dessutom reducerar den bruset i svansen. Men en av nackdelarna är att plotten inte ger en direkt visualisering av själva graden.

2.4.3 Klustring

Klustring innebär att det är en hög sannolikhet att två olika noder som har en gemensam granne också själva är sammanbundna, och om så är fallet får man en triangelstruktur. I verkligheten är sociala nätverk, liksom många andra typer av nätverk, klustrade, vilket innebär att nätverket har en relativ hög täthet av trianglar. I ett nätverk som beskriver vänskapsband kan detta förklaras med att det är en hög sannolikhet att din väns vän också är din vän.

Man kan kvantifiera graden av klustring i ett nätverk med *klustringskoefficienten* C, $0 \leq C \leq 1$, som mäter tätheten av trianglar i ett nätverk. Klustringskoefficienten kan definieras på två olika sätt. Ett sätt att kvantifiera klustringskoefficienten är som [10]

$$C = \frac{3 \times \text{antal trianglar i nätverket}}{\text{antal sammanbundna tripplar av noder}} \tag{5}$$

där en trippel är en mängd av tre noder. I själva verket så mäter C andelen av tripplar som har den tredje kanten som saknas ifylld och därmed bildar en triangel. Man kan säga att C är medelsannolikheten att två olika noder med en gemensam granne också själva är sammanbundna. Den alternativa definition av klustringskoefficienten kommer från Watts och Strogatz som föreslog en definition av ett lokalt värde som [10]

$$C_i = \frac{\text{antal trianglar kopplade till nod } i}{\text{antal tripplar centrerade på nod } i} \tag{6}$$

För noder med grad 0 eller 1 sätts $C_i = 0$. Sedan blir klustringskoefficienten för hela nätverket [10]

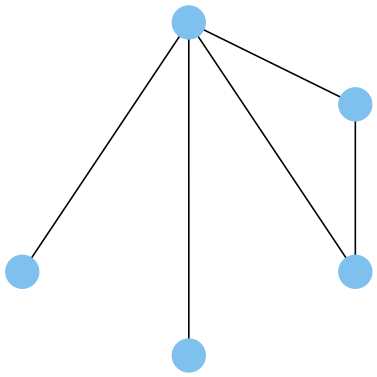
$$C = \frac{1}{N} \sum_i C_i \tag{7}$$

Definition (7) tenderar att ge mer vikt åt noder med låg grad, eftersom sådana noder har en liten nämnare i ekvationen och ger därför annorlunda resultat från definition (5). Skillnaden mellan de två illustreras nedan i figur 5 nedan.

De båda olika definitionerna av klustringskoefficienten ovan är empiriska mått. För att beräkna klustringskoefficienten för en slumpgraf kan man använda motsvarande väntevärden. Alternativt kan man beräkna den betingade sannolikheten att två noder är sammanbundna med en kant givet att de har en gemensam granne, och använda detta som mått på klustring.

Ett konkret exempel på ett nätverk med en hög klustring visas återigen i studien [5] gjort över samförfattarskap av artiklar inom neurologi publicerade mellan 1991 och 1998, som gav en klustringskoefficient (definierat enligt ekvation (7)) på $C = 0.76$.

Klustringskoefficienten mäter tätheten av trianglar av ett nätverk. En uppenbar generalisering skulle vara att studera tätheten av cykler med fyra eller fler noder, istället för tre, men tyvärr är denna teori ännu inte så välutvecklad.



Figur 5. Illustrering av hur C beräknas. Grafen har en triangel samt åtta sammanbundna tripplar av noder. Enligt ekvation (5) får man en klustringskoefficient $C = (3 \times 1)/8 = 3/8$. Enligt ekvation (6) får noderna de lokala klustringskoefficienterna: 1, 1, 1/6, 0, 0. Medelvärdet enligt ekvation (7) blir $C = 13/30$.

Nedan följer en tabell [10] över olika nyckeldata som beskrivits i detta kapitel, förutom gradkorrelationskoefficienten som beskrivs senare (kapitel 2.4.4.2).

Nätverk	N	m	z	l	γ	C^1	C^2	r
Samförfattarskap	253 339	496 489	3.92	7.57	-	0.15	0.34	0.120
Filmskådespelare	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208
Citationsnätverk	783 339	6 716 198	8.57		3.0/-			
WWW Altavista	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7			
Internet	10 697	31 992	5.98	3.31	2.5	0.035	0.39	-0.189
Kraftnät	4 941	6 594	2.67	18.99	-	0.10	0.080	-0.003
Neuralt nätverk	307	2 359	7.68	3.97	-	0.18	0.28	-0.226

Tabell 1. Nyckeldata för olika nätverk. Storheterna är :
N = totala antalet noder.
m = totala antalet kanter.
z = medelgrad.
l = medelnodavstånd.
Exponent γ för gradfördelningen om gradfördelningen följer en potenslag (eller "-" om ej, in/ut-grad ges om grafen är riktad).
C^1 = klustringskoefficient från ekvation (5).
C^2 = klustringskoefficient från ekvation (7).
r = gradkorrelationskoefficient.
Tomma rutor innebär att data saknas.

2.4.4 Andra egenskaper

De tre ovanstående egenskaperna är som sagt de viktigaste och som återkommer i de kommande kapitlen. Men för fullständighetens skull beskrivs i detta avsnitt ytterligare några egenskaper som observerats i verkliga nätverk och som är vanligt förekommande i litteraturen.

2.4.4.1 Assortativ blandning

Ett fenomen som kan ses i nätverk som består av flera olika nodtyper är att sannolikheten att en kant sammanbinder två olika noder beror på deras nodtyper. I sociala nätverk kallas den här typen av selektiva förbindelser för *assortativ blandning*. Assortativitet är således en nods förmåga att binda till andra noder som har samma eller andra egenskaper. Ett sådant fenomen finns i till exempel ekologiska nätverk. Ett ekologiskt nätverk kan bestå av olika arter som indelats i tre olika nodtyper beroende på om de är växter, växtätare och köttätare, och där kanterna representerar predator-byte relationer. Då ser man att många kanter förbinder växter med växtätare, liksom växtätare med köttätare, medan det är endast få kanter som förbinder köttätare med köttätare eller växter med köttätare.

Ett känt exempel på assortativ blandning är blandning av ras [10]. I en studie ingick det 1 958 par från staden San Francisco i Kalifornien. I studien datafördes av vilken ras samtliga individer var. Enligt tabell 2 som visar antalet par verkar individerna i studien företrädesvis välja sina partners från samma ras. Detta anses vara ett vanligt fenomen i många sociala nätverk – att man tenderar att umgås företrädesvis med människor som liknar oss själva på något sätt. Om raserna skulle representeras av fyra olika nodtyper skulle nästan varje nod ha den största sannolikheten att binda till en nod av samma typ. Om man vidare delar upp nodtyperna också på kön kan man kanske sedan göra ännu mer realistiska definitioner av de lokala egenskaperna.

		Kvinnor			
		Svart	Latinamerikan	Vit	Annat
Män	Svart	506	32	69	26
	Latinamerikan	23	308	114	38
	Vit	26	46	599	68
	Annat	10	14	47	32

Tabell 2. Par i studien sorterade efter individens ras. Rutor med samma färg visar siffror som jämför värden för män och kvinnor inom samma ras. Nästan lika många män som kvinnor av samma ras ingick i studien; kvoten mellan män och kvinnor är 1.1, 1.2 , 0.9 och 0.6 för svarta, latinamerikaner, vita respektive annat.

2.4.4.2 Gradkorrelation

Verkliga nätverk uppvisar *gradkorrelation* som innebär att det finns korrelationer mellan sammanbundna noders grader. Detta kan ses som ett specialfall av assortativ blandning där sannolikheten att två noder sammanbinds beror på vilken grad de har. Fenomenet kallas assortativ blandning, eller assortativitet, ifall noder med en hög grad oftare binder till andra noder med hög grad. I det andra fallet, om noder med en hög grad istället oftare binder till noder med en låg grad, kallas detta för disassortativ blandning, eller dissortativitet,

Korrelationen kan anges genom att beräkna korrelationskoeffcienten för nodernas grader i båda ändarna av en kant. Korrelationskoeffcienten bör vara positivt för assortativt blandade nätverk och negativa för disassortativa sådana. Verkliga nätverk har både assortativa och disassortativa blandningar, men det finns en intressant uppdelning efter nätverkstyp. Det är nämligen så att sociala nätverk tycks vara assortativa, medan de andra grupperna av nätverk (informationsnätverk, teknologiska nätverk, biologiska nätverk) verkar vara disassortativa. Detta kan ses i tabell 1 på korrelationskoeffcientens tecken.

2.4.4.3 Robusthet

En egenskap som kan vara av intresse är nätverkets *robusthet*, motståndskraft, vid borttagning av noder. När noder tas bort ökar avståndet mellan noderna, och om tillräckligt många noder tas bort börjar stigarna mellan noderna i nätverket att försvinna och nätverkets struktur sönderfaller allt mer. När man talar om ett nätverks robusthet menar man hur fort nätverket sönderfaller om man låter noder tas bort från nätverket.

Hur robust ett nätverk är beror inte bara på nätverkets struktur, utan också på vilket sätt noderna tas bort på. Det enklaste sättet som noderna kan tas bort på är att helt enkelt ta bort noder helt slumpmässigt ur hela grafen. Ett annat sätt är att inrikta sig på att ta bort noder av endast en viss typ, kanske alla de med den högsta graden. För de flesta nätverk gäller det att när hela tiden endast noder med den högsta graden tas bort, bryts nätverket ned snabbare än om noderna tas bort slumpmässigt. Detta beror på att när noderna med de högsta graderna tas bort, så tas samtidigt också bort flest kanter.

För sociala liksom andra typer av nätverk med en tungsvansad gradfördelning har numeriska simuleringar visat att nätverken är motsåndskraftiga mot borttagning av noder om detta sker på det helt slumpmässiga sättet, men att de däremot är lätta att bryta ned om man istället hela tiden tar bort noder med den högsta graden. I sådana nätverk har de flesta noder en låg grad, och endast en lite del av noder har en hög grad. När noderna tas bort helt slumpmässigt så är det en större sannolikhet att noder med en låg grad tas bort – vilket ju inte påverkar nätverket speciellt mycket. Om endast noder med den högsta graden tas bort försvinner som sagt många kanter med varje nod och nätverket tappar fort sin struktur.

Ett nätverks robusthet är särskilt viktigt inom epidemiologin, där borttagning av en nod i ett kontaktnätverk kan innebära att individen i fråga får en vaccination. Denna individ kan sedan tas bort från nätverket eftersom den inte längre kan få sjukdomen, men inte heller sprida den vidare. På så sätt skulle olika vaccinationsstrategier kunna ge olika effekter på smittspridningen.

2.4.4.4 Gruppstruktur

Ytterligare ett fenomen som de flesta sociala nätverk visar är *gruppstrukturer*. Detta innebär att i nätverket finns en uppdelnig av noder på grupper. Noder inom grupperna har en hög täthet av kanter inom gruppen, medan tätheten av kanterna mellan grupper är lägre. Det kan vara så att människor delar in sig i grupper efter intresse, yrke, ålder och så vidare, där man har många bekanta, medan endast en liten del av individerna i gruppen har bekanta över gruppgränser. Det är inte alltid lätt att hitta sådana gruppstrukturer. Men man skulle kunna tänka sig att man skulle kunna upptäcka gruppstrukturer i ett citationsnätverk, där grupperna är olika forskningsområden (se figur 3 b)). På samman sätt skulle ekologiska nätverk kunna delas in i delsystem inom ett ekosystem.

2.5 Kända studerade verkliga sociala nätverk

2.5.1 Erdőstal

Ett känt nätverk som studerats mycket är nätverket över samförfattandet av vetenskapliga artiklar bland matematiker. I detta nätverk är noderna matematiker och två matematiker sammanbinds med en kant om de har samförfattat en artikel. Vid studier av detta nätverk har man definierat *Erdőstalet* för en matematiker som är det antalet artiklar som en matematiker är från matematikern Paul Erdős, som var en mycket produktiv matematiker som skrev ungefär 1500 artiklar och har 511 medförfattare. Erdőstalet anger således ett samarbetsavstånd från en matematiker till Erdős.

För att tilldelas ett Erdőstal måste man samförfatta en artikel med en författare som själv har ett Erdőstal. Erdős själv har Erdőstalet 0. Han var som sagt mycket produktiv och har totalt 511 samförfattare, och alla dessa får Erdőstalet 1 (se tabell 3 nedan). Dessa medförfattare med Erdőstalet 1 ger ett Erdőstal 2 till andra författare genom att samförfatta en artikel med dessa (såvida de inte redan har ett lägre Erdőstal, i detta fall 1). På så sätt sprids Erdőstalet vidare till nya författare. En person utan någon som helst koppling till Erdős får ett oändligt, eller ett odefinierat, Erdőstal. På länken nedan kan man testa enskilda matematiker för att se hur långt de är från Erdős. Det är också möjligt att se avståndet mellan två matematiker.

`http://www.ams.org/msnmain/cgd/index.html`

Nätverket har ungefär 401 000 författare och ungefär 1.9 miljoner författade artik-

lar (juli 2004). De allra flesta av artiklarna har en eller ett fåtal författare. Ungefär 62.4% av artiklarna är författade av endast en författare, och 27.4% av artiklarna är författade av två författare. Detta gör att nästan 90 % av artiklarna är författade av färre än tre författare. Vidare gäller det att 8.0% av artiklarna är författade av tre författare, 1.7% av fyra författare, 0.4% av fem författare, och 0.1% av sex eller fler författare. Med tiden har antalet artiklar författade av endast en författare minskat stadigt över tiden. På 1940-talet var över 90% av artiklarna författade av endast en författare. För närvarande är motsvarande siffra under 50%. Hela grafen har ungefär 676 000 kanter, så det genomsnittliga antalet av medförfattare per person är 3.36.

`http://www.oakland.edu/enp`

Erdőstal	Antal matematiker
0	1
1	511
2	8162
3	33605
4	83642
5	87760
6	40014
7	11591
8	3146
9	819
10	244
11	68
12	23
13	5

Tabell 3. Tabell över Erdőstal.

2.5.2 Bacontal

I en intervju med filmskådespelaren Kevin Bacon kommenterade han att han har arbetat med alla i Hollywood, direkt eller via endast en skådespelare som länk. Detta medförde att man fick upp ögonen för vem som kunde vara den mest centrala skådespelaren i Hollywood, och begreppet *Bacontal* grundades. På samma sätt som man formulerat Erdőstalet i nätverket som beskrivits ovan formulerar man här ett Bacontal, vilket innebär att Bacontalet anger hur långt man är från Kevin Bacon i nätverket. Precis som för Erdőstalet gäller att endast Kevin Bacon själv har ett Bacontal på 0, och från detta utgår man när man beräknar Bacontal för övriga skådespelare.

Kevin Bacontal	Antal skådespelare
0	1
1	2251
2	225506
3	719767
4	178784
5	12205
6	1040
7	165
8	17

Tabell 4. Tabellen [7] visar Kevin Bacontal för samtliga skådespelare i nätverket.

I tabellen kan man se att på avståndet 0 från Kevin Bacon finns endast en skådespelare och det är Kevin Bacon själv. På avståndet 1 finns det 2251 skådespelare, alla de som medverkat i samma film som Kevin Bacon. På avståndet 2 finns det 225506 skådespelare, alla de som medverkat i en film tillsammans med en skådespelare som medverkat i en film med Kevin Bacon. Hur central en skådespelare är mäts med den genomsnittliga graden av åtskillnad (se kapitel 2.4.1), där en skådespelare är mer central ju lägre värde. Det genomsnittliga Kevin Bacontalet är 2.951023 och innebär att han inte är den mest centrala skådespelaren i Hollywood, utan kommer först på plats 507, men med tanke på nätverkets storlek på 1.6 miljoner är han ändå mycket mer central än de allra flesta skådespelare.

Mer information om detta projekt, där man listar de 1000 mest centrala skådespelarna (ändras med tiden), anger Bacontal samt avstånd mellan skådespelare, finns på länken

`http://www.cs.virginia.edu/oracle/`

Projektet har den mycket ambitiösa filmdatabasen Internet Movie Database

`http://www.imdb.com/`

att tillgå.

3 Erdős-Rényi modellen

Detta kapitel och de som följer handlar nu om slumpgrafmodeller för nätverk. Här inriktar man sig på att försöka skapa modeller för verkliga nätverk som kan användas för att generera nätverk som så mycket som möjligt liknar verkliga nätverk. Detta till skillnad från kapitel 2 som handlade om vilka egenskaper verkliga nätverk genom empiriska studier har visat sig ha. De viktigaste egenskaperna hos ett nätverk är som beskrivits i föregående kapitel en tungsvansad gradfördelning, korta avstånd mellan noderna, och en hög klustring. Detta är egenskaper som har observerats i sociala nätverk, men också i de andra typerna av nätverken.

Nu är målet alltså att formulera en modell som kan generera grafer med dessa egenskaper. Vissa av modellerna som beskrivs kommer att visa sig vara bättre lämpade än andra. En del modeller kanske har några egenskaper som stämmer överens med de verkliga nätverkens men brister när det gäller de andra egenskaperna, och bland modellerna kan det förhålla sig på så sätt att där den ena modellen brister där är den andra modellen lämplig, men där den ena modellen var lämplig är den andra modellen nu olämplig. Det ska också nämnas att det finns modeller som fångar upp alla de viktigaste egenskaperna, men tyvärr ändå genererar orealistiska grafer.

Även om modellerna först kommer att beskrivas finns redan här en sammanfattande tabell som visar vilka egenskaper hos respektive modell som överensstämmer med de verkliga nätverkens. Vad gäller bipartita modeller är det svårt att säga något generellt om deras egenskaper eftersom de finns i en stor variation. Men en bipartit modell definerad i kapitel 5.3 uppfyller både gradfördelningen och klustringen, men något resultat för medelnodavståndet finns inte tillgängligt. Tabellen är tänkt att fungera som en snabb sammanfattning som är enkel att hitta.

Modell	Gradfördelning	Medelnodavstånd	Klustring
Erdős-Rényi	Nej	OK	Nej
Small-world	Nej	OK	OK
Preferential attachment	OK	OK	Nej

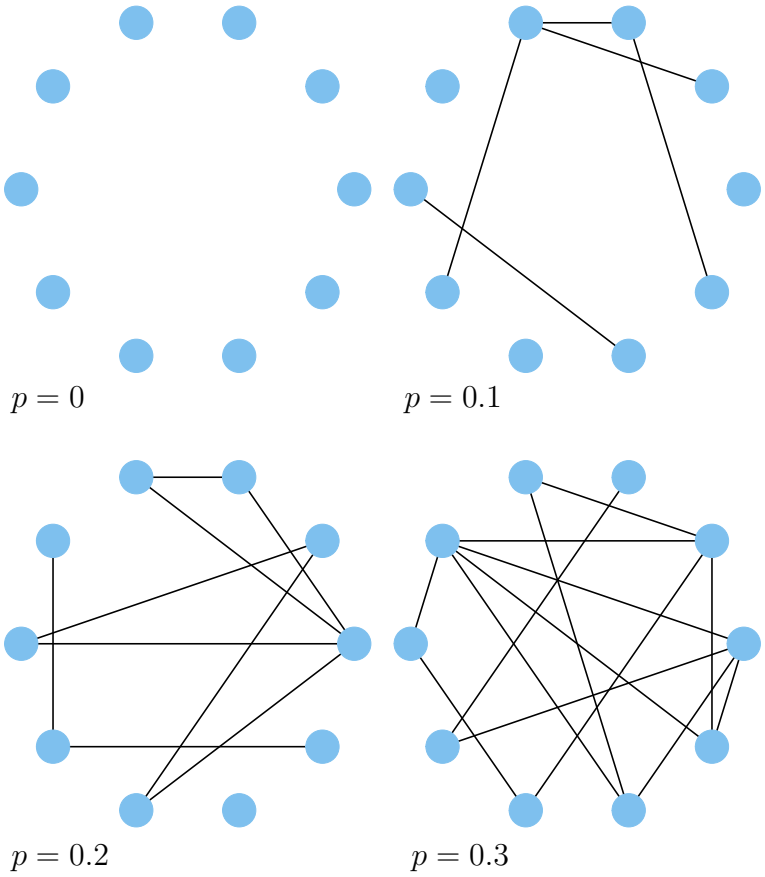
Tabell 5. En sammanfattande tabell över slumpgrafmodeller och vilka av deras egenskaper som stämmer överens med de verkliga nätverkens. Bipartita modeller finns inte med eftersom det är svårt att säga något generellt om deras egenskaper.

Det som beskrivs i detta kapitel är den enklaste slumpgrafmodellen för verkliga nätverk och dess egenskaper, men också varför den inte duger så bra som nätverksmodell, vilket kanske är typfallet när man gör avkall på komplexiteten. Även om denna modell kan betraktas som alltför enkel är den ändå viktig eftersom att den har viktiga egenskaper som dessutom kan beräknas exakt. Den är också den första slumpgrafmodellen, en klassiker alltså.

3.1 Erdős-Rényi grafen

Den första modellen för nätverk med en slumpmekanism definierades av Paul Erdős och Alfréd Rényi år 1959, efter att Erdős upptäckte att probabilistiska metoder ofta var användbara vid lösandet av problem inom grafteorin. Modellen kallas för *Erdős-Rényi grafen* eller *Poisson grafen.*

Det som i allmänhet karaktäriserar en slumpgraf är att den involverar någon typ av slumpmekanism. Erdős-Rényi grafen skapas genom att noderna är utsatta i förväg och kanterna läggs sedan till helt slumpmässigt, och oberoende av varandra, mellan noderna. Grafen betecknas $G_{N,p}$ där N är ett fixt antal noder och p är sannolikheten med vilken en kant uppstår mellan ett nodpar. Sannolikheten p är hela tiden konstant och samtliga kombinationer av nodpar testas innan grafen är klar. Hur många kanter som grafen får för ett visst N, det vill säga vad grafen får för kanttäthet beror på värdet på p. En grafs kanttäthet, eller dess *densitet* kan anges med ett densitetsmått som antar värden mellan 0 och 1 (se kapitel 2.1 för definitionen). Om $p = 0$ fås en graf med samtliga noder isolerade, och om $p = 1$ finns det en kant mellan samtliga nodpar. För $0 < p < 1$ fås ett resultat mellan dessa extremer, se figur 6 nedan för en illustrering av grafer som utvecklats med olika värden på p. Lägre värden på p ger en *gles* graf och högre värden på ger en *tät* graf.



Figur 6. Illustrering av hur olika värden på p påverkar Erdős-Rényi grafen utseende. $N = 10$ för samtliga grafer. Eftersom slumpen medför att även mer eller mindre osannolika utfall kan inträffa illustreras här endast de förväntade utfallen, det vill säga grafernas förväntade antal kanter Mp (se kapitel 3.2 nedan).

3.2 Erdős-Rényi grafens egenskaper

3.2.1 Gradfördelning

Av den ovan beskrivna processen med vilken grafen bildas inses att utvecklingen av en graf kan ses som ett slumpmässigt binomialförsök. Antalet kanter i hela grafen blir en stokastisk variabel som är $Bin(M, p)$, där M är *det totala maximala antalet kanter* som man kan få i grafen, som man får om det går en kant mellan samtliga nodpar,

$$M = \frac{N(N-1)}{2} \tag{8}$$

Det *förväntade antalet kanter* i hela grafen följer av väntevärdet för en binomialfördelning och är Mp.

Vilken *grad* en nod får blir också en binomialfördelad stokastisk variabel. Alltefter-som noden testas med samtliga andra noder blir den högsta möjliga grad som den kan få $N-1$ och den lägsta 0. En nods grad är binomialfördelad som $Bin(N-1, p)$, och har väntevärdet $p(N-1)$.

Sedan man börjat studera slumpgrafer har många egenskaper blivit exakt bestämda när antalet noder i grafen går mot oändligheten, det vill säga man har erhållit asymptotiska resultat när $N \to \infty$. En av dessa egenskaper är just gradfördelningen. När man beräknar gränsvärdet när N blir stort vill man att den förväntade graden för varje nod $\lambda = p(N-1)$ ska vara konstant. Detta gör man för att man inte vill att den förväntade graden ska öka bara för att man betraktar en större graf. Det skulle vara att ändra på de lokala egenskaperna som man sett empiriskt att nätverket har och som man nu också vill fånga upp i modellen. Att man vill behålla en konstant förväntad grad kan förklaras till exempel genom att betrakta ett socialt nätverk där kanterna är bekantskapsband mellan individer. Om man inte skulle hålla den förväntade graden konstant innebär detta att individerna i nätverket skulle ha fler bekanta bara för att de finns i en större population, vilket inte är fallet eftersom man har inte fler bekanta bara för att man till exempel bor i ett land med en större befolkningsmängd. När på detta sätt $N \to \infty$ innebär det att $p = \lambda/(N-1)$ så småningom blir mycket litet. Detta leder till att en nods grad som är binomialfördelad istället kan approximeras med en Poissonfördelning. Den asymptotiska sannolikheten p_k att en slumpmässigt vald nod har grad k fås alltså som

$$\binom{N}{k} p^k (1-p)^{N-k} \to \frac{\lambda^k e^{-\lambda}}{k!} := p_k \tag{9}$$

vilket är anledningen till att Erdős-Rényi grafer ibland kallas för Poissongrafer.

Att nodernas grader är Poissonfördelade innebär att de inte varierar så mycket eftersom Poissonfördelningens spridning inte är så stor. Detta är också en av Erdős-Rényi grafens nackdelar då många stora verkliga nätverk har en stor spridning av graderna, mycket större än Poissonfördelningens. Som beskrivits i kapitel 2 följer gradfördelningen hos nätverken ofta en potenslag med en lång högersvans, längre än Poissonfördelningens. Erdős-Rényi grafens gradfördelning stämmer alltså inte överens med

de verkliga nätverkens gradfördelning och bör därför inte heller användas till att försöka modellera verkliga nätverks tungsvansade gradfördelning. Anledningen till att grafen ändå är intressant är att den har andra viktiga egenskaper som dessutom kan beräknas exakt.

3.2.2 Klustring

Inte heller när det gäller klustring lyckas Erdős-Rényi grafen särskilt bra med att avbilda verkliga nätverk, eftersom den har en för låg klustring. Klustring innebär att det förekommer ovanligt många trianglar i grafen, vilket har sin grund i att det är en förhöjd sannolikhet att två noder som sammanbinds till en gemensam nod också själva är sammanbundna. Men i definitionen av hur Erdős-Rényi grafen skapas är $p = \lambda/(N-1)$ konstant och kanterna adderas oberoende, vilket innebär att sannolikheten att två noder som sammanbinds till en gemensam nod också själva är sammanbundna också är p, det vill säga det förekommer ingen förhöjd kantsannolikhet i modellen. Vidare gäller det också att klustringen går mot 0 när nätverket blir stort.

3.2.3 Erdős-Rényi grafens fasövergång

En egenskap som Erdős-Rényi grafen har är en *fasövergång*. Som namnet antyder så sker det en förändring av något slag. Det är den realiserade grafens struktur som förändras med värdet på λ. Kantsannolikheten p styr som ovan beskrivits grafens kanttäthet, där ett litet λ skapar en gles graf och ett stort λ skapar en tätare graf. Men därtill styr λ också grafen struktur, och beroende på om λ är stort eller litet skapas en graf som tillhör en av de två möjliga faserna.

Vid ett litet λ skapas en graf med ett fåtal kanter och där alla komponenter är små med en exponentiell storleksfördelning och en ändlig förväntad storlek. Den största komponenten blir i detta fall av ordningen $\ln N$. I det andra fallet när λ är stort får grafen en struktur där de flesta noder tillhör en och samma jättestora, *gigantiska* komponent, som är en komponent av samma storleksordning som hela grafen. De noder som inte ingår i den gigantiska komponenten bildar små komponenter med precis som tidigare en exponentiell storleksfördelning och med en ändlig förväntad storlek (se figur 7 och 8). Denna fasövergången inträffar när $\lambda = 1$.

Den förväntade storleken av den gigantiska komponenten, som uppstår när $\lambda > 1$, kan beräknas med en formel (ekvation (11)) som kan härledas via följande heuristiska resonemang. Om u är andelen av noderna som inte tillhör den gigantiska komponenten, kan u också tolkas som den sannolikhet att en slumpmässigt vald nod ur grafen inte tillhör den gigantiska komponenten. Denna sannolikhet är densamma för alla noder som inte tillhör den gigantiska komponenten. Detta innebär att om en nod har k grannar blir sannolikheten att ingen av dem tillhör den gigantiska komponenten u^k. Om man viktar detta uttryck med sannolikhetsfördelningen för k, se

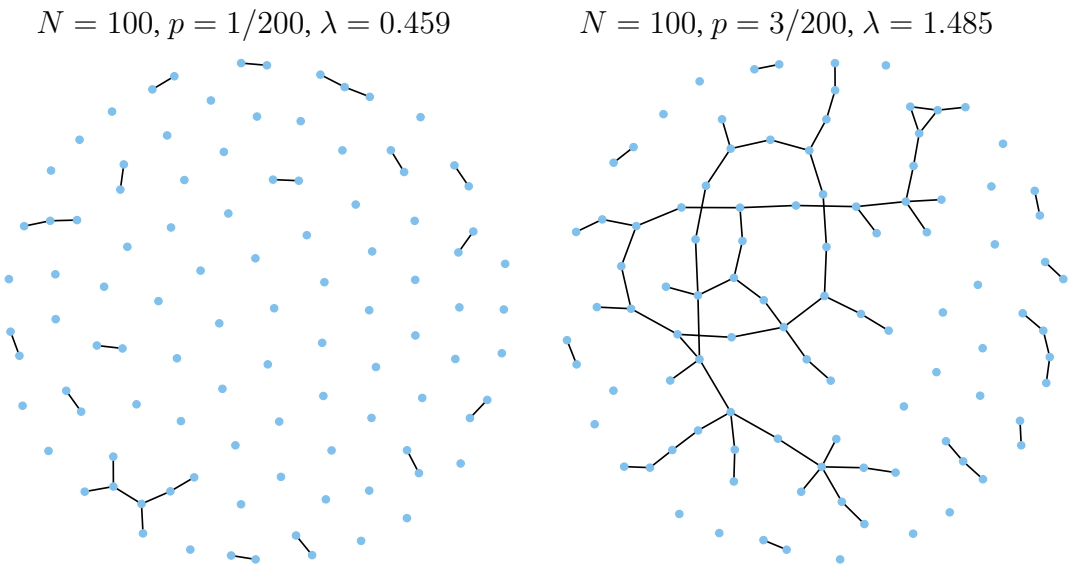
ekvation (9) ovan, får man följande relation för u när $N \to \infty$ [10]

$$u = \sum_{k=0}^{\infty} p_k u^k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda u)^k}{k!} = e^{\lambda(u-1)} \tag{10}$$

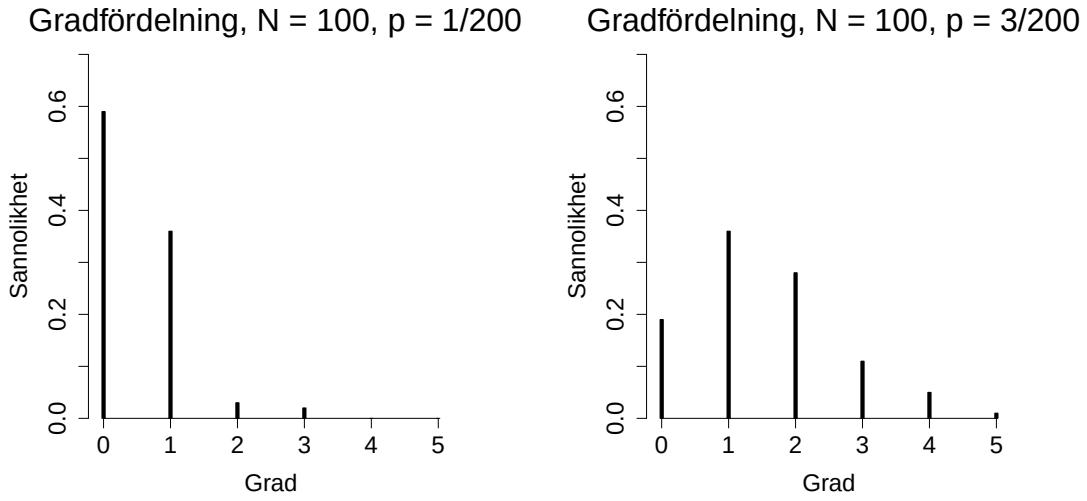
Låt S vara andelen av noderna som tillhör den gigantiska komponenten, vilket också är sannolikheten att en slumpmässigt vald nod tillhör den gigantiska komponenten. Detta innebär att $S = 1 - u$ och S är lösningen till [10]

$$S = 1 - e^{-\lambda S} \tag{11}$$

Från ekvation (11) kan man se att om $\lambda < 1$ är ekvationens enda icke-negativa lösning $S = 0$. För $\lambda > 1$ finns det också en icke-negativa lösning, som är just den gigantiska komponentens storlek.



Figur 7. Illustrering av hur olika värden på p ger olika grafer. Endast förväntade utfall illustreras. *Till vänster* : eftersom $\lambda = 0.459 < 1$ finns ingen gigantisk komponent och alla andra komponenter är små. *Till höger* : $\lambda = 1.485 > 1$ och man kan se en gigantisk komponent och även här är alla andra komponenter små.



Figur 8. Gradfördelningen för graferna i figur 7. *Till vänster*: gradfördelningen för grafen utan den gigantiska komponenten. *Till höger*: gradfördelningen för grafen med den gigantiska komponenten. Kanske kan man också se i histogrammen att gradfördelningen är approximativt Poissonfördelad.

3.2.4 Medelnodavstånd

Avståndet mellan noderna i grafen växer logaritmiskt med N. För $\lambda < 1$ är detta uppenbart eftersom den största komponenten i grafen då är av storleksordning $\ln N$. För $\lambda > 1$ innehåller grafen en gigantisk komponent av storleksordning N och man skulle alltså kunna tänka sig att det var långt mellan noderna i denna komponent. Det går dock att visa att avståndet mellan två noder i den gigantiska komponenten endast växer logaritmiskt med N. Mer precist så gäller att avståndet mellan två noder i den gigantiska komponenten är högst $c \ln N$, för någon konstant c, med en sannolikhet som går mot 1 då $N \to \infty$. Med andra ord uppfylls villkoret (2) för "världen är liten"-fenomenet av Erdős-Rényi grafen. I [3] ges exempel som visar att medelnodavståndet för Erdős-Rényi grafen är nära medelnodavståndet för verkliga nätverk av samma storlek.

3.3 Generaliseringar av Erdős-Rényi grafen

3.3.1 Konfigurationsmodellen

Eftersom Erdős-Rényi grafen inte fångar upp de verkliga nätverkens gradfördelning behöver man en annan modell för att göra detta. Fler modeller kommer i de följande kapitlen, men det finns också en möjlighet att generalisera Erdős-Rényi grafen genom att skapa en modell med en i förväg bestämd gradfördelning, som till exempel en potenslag. En sådan modell, som skapas genom en algoritm som följer nedan, kallas för *konfigurationsmodellen*.

Modellen skapas genom att p_k definieras som den önskade andelen noder i grafen med grad k. Sedan väljs en *gradsekvens*, som är en mängd av N tal, $\{k_1, ..., k_N\}$, där varje tal k_i, $1 \leq i \leq N$, anger graden för nod i, $i = 1, ..., N$. Mer precist så dras gradsekvensen från fördelningen p_k under vilkoret att summan $\sum_i k_i$ är jämn. Om man tänker sig att k_i för nod i är antalet halvkanter som sticker ut från noden, väljs sedan likformigt slumpmässigt par av sådana halvkanter som sammanbinds. Detta sätt att konstruera modellen på innebär att nätverket som skapas får den önskade gradfördelningen. Observera att denna model tillåter multipla kanter (två eller fler kanter som sammanbinder ett nodpar) och loopar (kanter som går från en nod tillbaka till sig själv).

4 Small-world modeller

Som beskrivits i kapitel 3 så duger inte Erdős-Rényi grafen till att modellera verkliga nätverk eftersom den bland annat har en för låg klustring. Den egenskap som grafen dock har och som överensstämmer med det som framkommit om verkliga nätverkens egenskaper är "världen är liten"-fenomenet, vilket innebär att avstånden mellan noderna är korta. Grafen är ju som bekant endast den enklaste modellen och som kan vidareutvecklas till en mer realistisk modell. Därför har man försökt utveckla Erdős-Rényi grafen så att den får en högre klustring och därmed lämpar sig bättre för modelleringen.

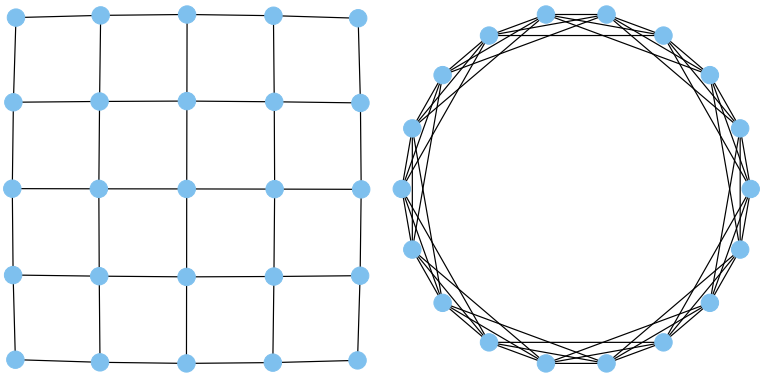
Duncan J. Watts och Steven Strogatz föreslog 1998 en *small-world* modell [12] som är just sådan att den har behållit Erdős-Rényi grafens korta nodavstånd men samtidigt också har en högre klustring än Erdős-Rényi grafen.

Det finns lite olika varianter på small-world modeller. Den ursprungliga modellen är ändå den som Watts och Strogatz har föreslagit. Utifrån den kan man sedan göra vissa modifieringar. I detta kapitel tas två small-world modeller upp och deras egenskaper beskrivs. Den ena är den ursprungliga modellen som föreslagits av Watts och Strogatz, *WS-modellen*, och den andra är en modell som föreslagits oberoende av Monasson och av Newman och Watts. Denna är en förenkling av WS-modellen som är enklare att analysera.

4.1 Small-world modeller

Small-world modeller skapas genom att man från början har, precis som i fallet med Erdős-Rényi grafen, N fixa, i förväg utsatta noder. Men denna gång slumpas inte kanterna ut, utan samtliga kanter är precis som noderna också utsatta i förväg och dessutom också på ett mycket bestämt sätt. Det är nämligen så att hela grafen från början är ett lågdimensionell regelbundet gitter. Small-world modeller kan byggas på gitter av vilken dimension eller topologi som helst, men det i särklass mest studerade fallet är den endimensionella.

Ett *gitter*, se figur 9 nedan, kan ses som en helt regelbunden graf med ordnade noder och kanter. Det är enkelt att se att klustringen i de båda gittrarna är hög samt att en nods klustringskoefficient beror endast på positionen i gittret och att många noder får ett precis likadant värde. Att det endast är positionen av en nod som påverkar dess klustringskoefficient innebär att klustringskoefficienten är oberoende av grafstorleken N. En annan viktig sak att lägga märke till är att medelavståndet mellan noderna i de flesta fall blir stort om N är stort.

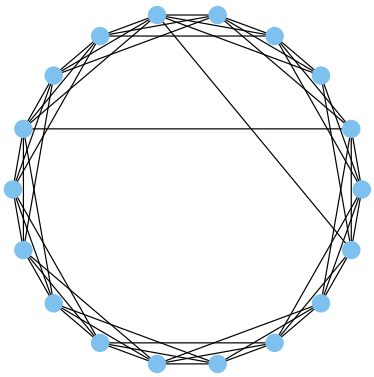


a) b)

Figur 9. Illustrering av två regelbundna gitter. Båda gitterna har en hög klustring samt ett stort medelnodavstånd. Gitter b) som används för att modellera small-world modeller har en ringstruktur där N noder är ordnade i en ring i vilken dem sammanbinds till samtliga andra noder som är $k = 3$ eller färre kanter ifrån. Totala antalet kanter i gitter b) är Nk.

Det vanligaste gittret som man tänker sig och som illustreras i figur 9 a) är ett tvådimensionellt kvadratiskt men de som används i small-world modeller är istället endimensionella och ordnade i en ring, se figur 9 b). Denna ringstruktur är uppbyggd på ett sådant sätt att noderna är ordnade i en ring och sammanbinds med alla noder inom avstånd k åt varje håll i denna ring. I allmänhet antas här att k är mycket mindre än N. I fortsättningen förutsätts att gittern alltid har ringstrukturen. Klustringen i denna struktur är som sagt hög, vilket är en egenskap som Erdős-Rényi grafen inte har. Men samtidigt är medelnodasvtåndet stort, det vill säga "världen är liten"-fenomenet finns ännu inte. Det återstår nämligen en sak till innan modellen är klar.

Small-world grafen skapas nu genom att med en sannolikhet p slumpmässigt förflytta kanterna inom gittret, se figur 10 nedan. Detta förflyttande gör att det skapas genvägar i nätverket vilket gör att avstånden mellan noder sjunker drastiskt.



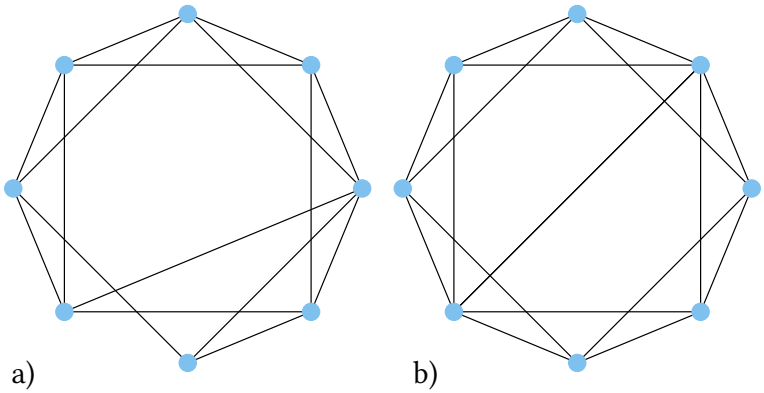
Figur 10. Illustrering av på vilket sätt WS-modellen minskar nodavstånden och skapar "världen är liten"-effekten, nämligen genom att förbinda avlägsna delar i nätverket.

Förflyttningen sker, i alla fall för WS-modellen, genom att varje kant beaktas i tur och ordning och med sannolikheten p förflyttas en ände av kanten till en ny slumpmässigt vald nod i ringen, med restriktionen att inga dubbla kanter eller loopar får bildas. Detta resulterar i att endast en liten andel av kanterna kommer att förflyttas och på detta sätt binder samman avlägsna delar i grafen så att nätverket får den eftersträvade "världen är liten"-effekten. Tätheten av dessa kanter som förbinder avlägsna delar blir låg och klustringen förblir hög, men samtidigt minskas medelnodavståndet mycket. Detta innebär att denna process tillåter grafen att skifta mellan ett helt ordnat gitter till nästan en Erdős-Rényi graf, beroende på valet av p.

Om sannolikheten p med vilken förflyttningen av kanterna görs är $p = 0$ fås ett helt regelbundet gitter (figur 9 b)) där varje nod binder till samtliga noder inom avstånd k i ringen. Det totala antalet kanter blir Nk. Klustringskoefficienten blir i detta fall $C = (3k - 3)/(4k - 2)$ och går mot $3/4$ för stora k. Att klustringskoefficienten beräknas enligt ovanstående formel kan förstås genom att använda sig av ekvationerna (6) och (7) på följande sätt. Gitterns struktur gör att alla noder har samma klustringskoefficient, vilket innebär att $C = C_i$. Vidare kan man se att klustringskoefficienten bara kommer att bero på k, där $k \geq 1$. För $k = 1$ ingår inte noden i någon triangel och täljaren i ekvation (6) blir 0, liksom hela den lokala klustringskoefficienten. För $k = 2$ ingår noden i 3 trianglar, för $k = 3$ ingår noden i 9 trianglar, och så vidare på grund av en systematisk utveckling när k växer. Detta innebär att täljaren ska anta värden $0, 3, 9, 18, 30, ...$ för $k = 1, 2, 3, 4, 5, ...$, och just en sådan följd ges av funktionen $3(k^2 - k)/2$. För $k = 2$ är 6 tripplar centrerade på noden, för $k = 3$ är 15 tripplar centrerade på noden, och så vidare. Detta innebär att nämnaren i ekvation (6) ska anta värden $1, 6, 15, 28, 45, ...$ för $k = 1, 2, 3, 4, 5, ...$, och en sådan följd ges av funktionen $2k^2 - k$. Dessa resultat ger att ekvation (6) kan skrivas som $C = (3(k^2 - k)/2)/2k^2 - k$ som efter förenkling fås det ovan givna resultatet $C = (3k - 3)/(4k - 2)$.

När $p = 1$, förflyttas varje kant till en ny slumpmässigt vald nod och grafen blir nästan en Erdős-Rényi graf med typiska nodavstånd av ordningen $\ln N$, och klustringen blir mycket låg $C \sim 2k/N$. Det finns dock ett ganska stort område mellan dessa två extremer ($p = 0$ och $p = 1$) där modellen har både korta nodavstånd och hög klustring (kapitel 4.2.2). Den största begränsningen hos WS-modellen är att grafen som skapas får en gradfördelning med en för liten spridning. Vidare har modellen ett fixt antal noder och kan därför inte användas för att modellera nätverk som växer (se kapitel 6).

En variant av den ursprungliga Watts och Strogatz small-world modellen som förenklar analysen är en modell där man vid förflyttning av en kant flyttar båda kantens ändar och inte bara en, och genom att tillåta både dubbla kanter och loopar.



Figur 11. Illustrering av small-world modeller. a) WS-modellen. b) En variant av Watts och Strogatz där kanterna adderas. I figuren har endast en kant adderats.

Ytterligare en annan variant är en modell som föreslagits oberoende av Monasson och av Newman och Watts. I denna modell förflyttas inga kanter utan fler kanter adderas istället till gittern och förbinder slumpmässigt valda noder, se figur 11 b) ovan. Sannolikheten p definieras som sannolikheten per kant som finns i gittret att det adderas en ny kant någonstans i hela gittret. Detta gör att det förväntade antalet adderade kanter är Nkp och den förväntade graden är $2k(1+p)$. Denna modell har den eftertraktade egenskapen att inga noder blir avskilda från resten av nätverket, och detta gör att medelnodavståndet alltid är ändligt.

4.2 Small-worlds modellernas egenskaper

4.2.1 Medelnodavstånd och klustring

Av small-world modellernas uppbyggnad kan man se att det finns ett nära samband mellan medelnodavstånd och klustring. När man vill göra modellen mer realistisk med tanke på medelnodavståndet och inför flera förbindande kanter får man på samma gång en lägre klustring. På liknande sätt ger ett färre antal kanter en högre klustring men samtidigt också ett längre medelnodavstånd. För att modellen ska vara realistisk vill man gärna ha både korta nodavstånd och hög klustring. I simuleringar [12] har man faktiskt sett att för WS-modellen finns det en region för p mellan 0 och 1 som ger modellen samtidigt en hög klustring och ett litet medelnodavstånd, och modellen således får båda de eftertraktade egenskaperna (se figur 12 nedan). Att en sådan region existerar beror på att i början är klustringen hög och den ena egenskapen för modellen är redan uppfylld. När kanterna senare skapar genvägar i modellen kortas avstånden mellan noderna drastiskt, medan klustringen inte påverkas lika mycket till en början. På så sätt uppfylls de båda egenskaperna.

Medelnodavståndet och klustringen är de egenskaper i small-world modeller som fått mest uppmärksamhet, men trots det har man tyvärr ännu inte funnit någon exakt lösning för medelnodavståndet, utan endast några delresultat. För WS-modellen

vet man dock att när $p \to 0$ går medelnodavståndet mot $l = N/4k$, och ju större p är desto mer liknar modellen en Erdős-Rényi graf.

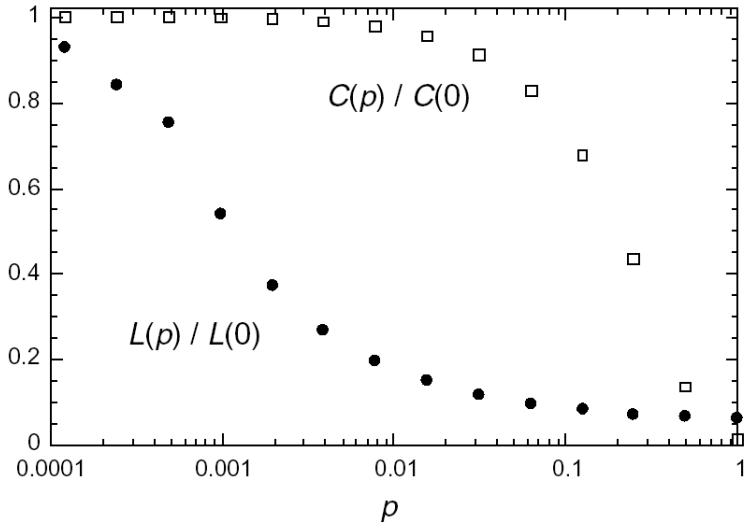
Nedan anges utan härledning två formler för klustringskoefficienten, tagna från [10]. Klustringskoefficienten för WS-modellen ges av

$$C = \frac{3(k-1)}{2(2k-1)}(1-p)^3 \tag{12}$$

För den modifierade modellen där kanterna adderas blir motsvarande

$$C = \frac{3(k-1)}{2(2k-1)+4kp(p+2)} \tag{13}$$

Av de båda formlerna ser man att klustringen inte beror av N och alltså inte går mot 0 i stora grafer.



Figur 12. Från en simulering [12] framkommer det tydligt att det finns en region för p mellan 0 och 1 som ger WS-modellen samtidigt en hög klustring och ett litet medelnodavstånd. I bilden illustreras klustringskoefficienten $C(p)$ och medelnodavståndet $l(p)$ som funktion av p i WS-modellen. $C(p)$ och $l(p)$ har normailserats med $C(0)$ respektive $l(0)$. En logaritmisk skala har använts för p för att bättre kunna se när medelnodavståndet minskar. Man ser att för låga medelnodavstånd kan klustringskoefficienten fortfarande vara hög, och således uppfylls de båda egenskaperna samtidigt.

4.2.2 Gradfördelning

Small-world modellernas gradfördelning överensstämmer inte med de flesta verkliga nätverkens. För den modifierade modellen där kanterna adderas har varje nod minst grad $2k$ för gittern som är från början, samt ett binomialt fördelat antal adderade kanter. Detta ger att sannolikheten p_j att ha grad j är [10]

$$p_j = \binom{N}{j-2k} \left[\frac{2kp}{N}\right]^{j-2k} \left[1 - \frac{2kp}{N}\right]^{N-j+2k} \tag{14}$$

för $j \geq 2k$, och $p_j = 0$ för $j < 2k$.
För WS-modellen förflyttas kantens ändar från noderna på ett sådant sätt att en nods lägsta grad kan endast bli k, och motsvarande sannolikhet blir i detta fall [10]

$$p_j = \sum_{n=0}^{min(j-k,k)} \binom{k}{n} (1-p)^n p^{k-n} \frac{(pk)^{j-k-n}}{(j-k-n)!} e^{-pk} \tag{15}$$

för $j \geq k$, och $p_j = 0$ för $j < k$.

Ingen av gradfördelningarna ovan beter sig som en potenslag. Av metoderna med vilka modellerna skapas inser man att gradfördelningen inte får en stor spridning med några noder med en grad långt över den förväntade graden, som en potenslagsfördelning skulle innebära. Från början har ju alla noder samma grad och det antal kanter som sedan tillkommer eller dras bort från en given nod har en fördelning av binomialtyp. Detta kan inte leda till en tungsvansad gradfördelning.

5 Bipartita grafmodeller

Detta kapitel handlar om bipartita grafer och precis som de tidigare graferna som tagits upp kan också denna typ av grafer användas till att modellera olika nätverk. Den bipartita grafen används till att modellera nätverk som kallas för *tillhörighetsnätverk*, och innebär i korthet ett nätverk där det finns en koppling mellan, säg två individer, om de har något gemensamt. Detta innebär att en bipartit graf illustrerar individer som har något gemensamt, som att ha medverkat i samma film till exempel. Man kan säga att i en bipartit graf delas individer in i grupper där alla individer i en grupp har en sak gemensamt.

Eftersom det finns många bipartita modeller studeras här endast en bipartit modell, som definieras i kapitel 5.3.

5.1 Tillhörighetsnätverk

Det finns olika sociala nätverk där en bipartit grafmodell kan vara lämplig. Till skillnad från de tidigare modellerna så har en bipartit modell två olika typer av noder och endast noder av olika typer kan sammanbindas, vilket innebär att analysen av grafen också blir annorlunda mot tidigare. Just det faktum att man valt att ha med två olika typer av noder tyder på att även nätverken som man vill studera skiljer sig från de tidigare.

Nätverk som modelleras med bipartita grafer brukar kallas för tillhörighetsnätverk, på engelska *affiliation networks*, och skiljer sig alltså från de tidigare. Exempel på tillhörighetsnätverk är nätverk av verkställande direktörer, styrelser, samt samarbeten av olika slag, som till exempel samarbeten mellan forskare eller filmskådespelare.

I exemplet med samarbeten mellan filmskådespelare, där ett samarbete föreligger mellan två skådespelare om de medverkat i samma film (kapitel 2.5.2 Bacontal), ser modelleringen av nätverket ut på följande sätt att den ena typen av noder är olika filmer och den andra typen är olika skådespelare. Kanterna i nätverket går endast mellan de olika nodtyperna, vilket i detta fall betyder att skådespelare bara kan sammanbindas med filmer, och tvärtom att filmer bara kan sammanbindas med skådespelare. Detta innebär att man får en bipartit skådespelare-film graf som visar vilka skådespelare som medverkat i en viss film men också i vilka filmer en viss skådespelare medverkat i. I exemplet med samarbeten mellan forskare (kapitel 2.5.1 Erdőstal) som definieras som att ett samarbete föreligger om de båda stått som medförfattare till en artikel, är forskarna och artiklarna de två olika nodtyperna. Kanterna i grafen går endast mellan författare och artiklar. Grafen visar således vem som skrivit vilken artikel och vilka författarna är till varje artikel. Ett annat exempel på en bipartit graf i en annan huvudgrupp är metaboliska nätverk, där nodtyperna kan vara substrat och reaktioner.

Det utmärkande för dessa tillhörighetsnätverk är att en koppling mellan individerna definieras på basis av något annat, till exempel så finns det en koppling mellan film-

skådespelare om de medverkat i samma *film*, och för att det ska finnas en koppling mellan forskare gäller det att de båda stått som medförfattare till en viss *artikel*, det är alltså filmen respektive artikeln som skapar en koppling mellan individerna och inte som det hittillls varit att två individer har en koppling om de har ett direkt samband som till exempel ett vänskaps- eller bekantskapsband. Men sociala näverk definieras utifrån relationer, och att ha något gemensamt kan i en vidare bemärkelse också ses som en relation.

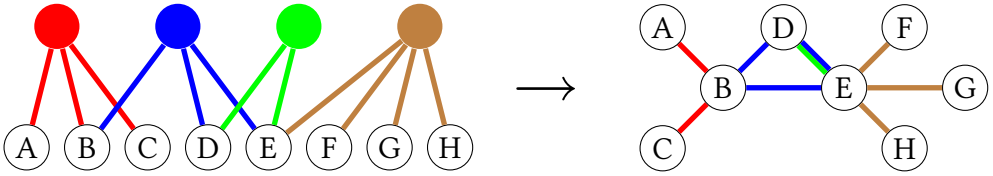
Detta innebär att man kan säga att man här studerar grupper, istället för som tidigare där man fokuserat på varje individ och kopplingar mellan dem. Här kanske man till exempel vill lista den grupp av skådespelare som varit med i en viss film eller också tvärtom för en skådespelare lista den grupp av filmer som denne medverkat i.

Metoder för att studera bipartita nätverk är mindre utvecklade än metoder för studier av enpartita nätverk, det vill säga de vanliga nätverken där det bara finns en nodtyp, och därför fnns det inte heller lika många resultat för dem. Men från ett bipartit nätverk kan man alltid få fram ett enpartit nätverk genom att göra en projektion på bara den ena nodmängden, se kapitel 5.2 nedan. I många fall där graferna är bipartita studeras de faktiskt först efter att ha projicerat dem på endast en mängd noder. Till exempel i kapitel 2.5 om Erdőstal och Bacontal behandlas nätverken som enpartita trots att de skulle kunna behandlas som bipartita enligt beskrivningen ovan. Figur 3 a) illustrerar nätverk över samförfattandet av vetenskapliga artiklar med endast en nodtyp.

5.2 Bipartita grafen

Den bipartita grafen kan skrivas $G = (U, V, E)$ där U och V är två disjunkta nodmängder, $U \cap V = \emptyset$, och E är kantmängden där kanterna endast får gå mellan två noder som inte tillhör samma mängd. Definitionen ovan kan uttryckas så att en bipartit graf är en graf med två olika typer av noder och där kanter kan endast gå mellan två noder av olika typer, se figur 13 a). Om U och V har samma antal element kallas G för en balanserad bipartit graf.

Från en bipartit graf kan man enkelt konstruera en enpartit graf genom att projicera den bipartita grafen på den ena nodmängden, se figur 13. En projektion av den bipartita grafen $G = (U, V, E)$ ger den enpartita grafen $G' = (V, E')$ där två noder $\{V_i, V_j\}$ i mängden V återfinns i E' om och endast om det finns en nod i U som de båda är kopplade till.



a) b)

Figur 13. En bipartit graf (a) och dess V-projektion (b). Kanten mellan noderna D och E har erhållits två gånger eftersom både nod D och E har två gemensamma grannar.

5.3 En modell för ett tillhörighetsnätverk

Det finns många tänkbara modeller för bipartita grafer. Här beskrivs endast en modell vars gradfördelning och klustring stämmer överens med de verkliga nätverkens. Vad gäller medelnodavståndet finns det tyvärr inget resultat ännu.

En bipartit graf [6] kan skapas genom att låta V beteckna en mängd av N noder och U en mängd av N' noder. För att grafen ska få en bra struktur väljer man $N' = \lceil \beta N \rceil$ för någon konstant $\beta > 0$. Till varje nod V_i i V ges oberoende en stokastisk vikt W_i enligt någon fördelning F som antas ha väntevärde 1. En bipartit graf skapas genom att med sannolikhet p_i oberoende sammanbinda nod V_i till noderna i U (ett försök görs alltså för varje nod i U), där

$$p_i = \gamma W_i n^{-1} \qquad i = 1, 2, ..., n \tag{16}$$

för någon konstant $\gamma > 0$. Grafen G som man är intresserad av fås sedan som beskrivits i föregående avsnitt genom att projicera på nodmängden V, det vill säga två noder i V sammanbinds om de båda är kopplade till en och samma nod i U. Om man tänker på noderna i V som individer och noderna i U som grupper representerar G alltså en graf där två indivier har en länk mellan sig om de är medlemmar i samma grupp. Vikterna $\{W_i\}$ på individerna kan man tänka på som sociala index där en individ med ett högre index har en större tendens att gå med i grupper och därigenom skapa kontakter med andra individer.

Graden D_i för nod V_i i grafen G kommer att vara fördelad som en summa av ett Poisson(βW_i) antal Poisson(γ)-variabler. Det blir på detta sätt eftersom antalet grupper som individ i är medlem i är binomialfördelat med parametrar m och p_i och alltså asymptotiskt Poisson(βW_i)-fördelat. Vidare fås storleken av en given grupp som summan av n stycken kantindikatorer, en för varje individ, vilket antyder att den är asymptotiskt Poisson-fördelad. Väntevärdet är $\sum_{i=1}^{N} E[p_i] = \gamma$ eftersom $E[W_i] = 1$, det vill säga man får en Poisson(γ)-fördelning. Det går att visa att fördelningen för D_i följer en potenslag om vikterna $\{W_i\}$ följer en potenslag och att exponenten i potensfördelningen blir densamma som exponenten i fördelningen för vikterna.

När det gäller klustringen så kan man visa att betingat på vikterna så är sannolikheten att två noder i och j som båda är för sammanbundna med nod k i grafen G

också är förbundna med varandra asymptotiskt är $(1 + \beta\gamma W_k)^{-1}$. Klustringen går alltså inte mot 0 då antalet noder går mot oändligheten och går att variera genom att ändra på modellens parametrar. Detta gör att klustringen överensstämmer med de verkliga nätverkens. Att klustringen blir som den blir beror på att det förväntade antalet grupper som individ k är medlem i är $\beta\gamma W_k$. Om detta är litet är det troligt att kanterna till i och j har uppkommit via samma grupp, det vill säga klustringen är stor eftersom det då också måste finns en kant mellan i och j. Om $\beta\gamma W_k$ å andra sidan är stort, blir det mindre troligt att kanterna till i och j har uppkommit via samma grupp och klustringen minskar därigenom.

6 Preferential attachment modeller

En naturlig egenskap som många verkliga nätverk har, och som endast nämnts i förbigående i kapitel 2, är att de faktiskt växer med tiden i och med att ny information tillkommer. Till exempel i nätverket över filmskådespelare i kapitel 2.5.2 växer nätverket hela tiden i och med att nya filmer produceras. Nätverket över webben växer när nya internetsidor uppstår och får en länk till andra internetsidor. Men exemplen är förstås många. Ett växande nätverk modelleras lämpligen med en graf som själv har en förmåga att växa. Genom att nya noder hela tiden introduceras in i grafen och sammanbinds till existerande noder fås en växande graf. Det finns olika sådana *regenerativa* modeller, och utvecklingar av dessa, för växande nätverk.

I detta kapitel beskrivs en regenerativ *preferential attachment* modell. Preferential attachment syftar till mekanismen med vilken de nya noderna sammanbinds med de existerande. Här beskrivs endast en preferential attachment mekanism, men det finns flera definierade. Egentligen har dessa mekanismer en lång historia, men det var Albert-László Barabási och Réka Albert som återupptäckte dem 1999, vilket sedan ledde till att populariteten för modeller med en potenslagfördelning blev så som den är idag.

6.1 Modellering med preferential attachment

Många, kanske de allra flesta verkliga nätverk växer med tiden. På grund av denna tillväxt finner man det som sagt lämpligt att också använda sig av en modell där antalet noder växer med tiden. Man försöker alltså här finna en modell som både växer och som vanligt också har andra egenskaper som överensstämmer med de verkliga nätverkens. När det gäller gradfördelningen är det lämpligt om den följer en potenslag, precis som för de tidigare beskrivna nätverksmodellerna. En graf som växer skapas enkelt genom att hela tiden, en åt gången, introducera och inkludera nya noder in i grafen. Det är önskvärt att sammanbinda noderna på ett sådant sätt att grafen får en potenslag som gradfördelning. För att ta redan på hur noderna ska inkluderas kan man kanske få några ledtrådar genom att observera ett verkligt nätverk och se vilka fenomen som styr när en ny nod (till exempel en individ) inkluderas. Om den simulerade modellen sedan visar sig att inte stämma tillräckligt bra överens med verkligheten kan det bero på att man inte beaktat vissa fenomen.

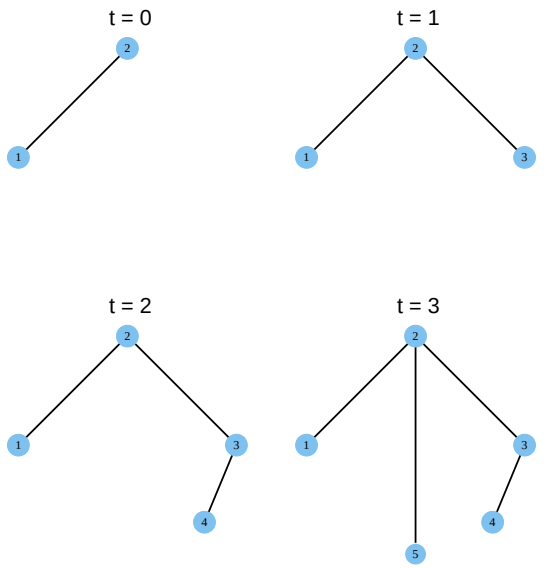
Introduceringen av nya noder har man löst på följande sätt att noder introduceras in i grafen och sammanbinds med redan existerande med en *sannolikhet som är proportionell mot deras grad*. Detta innebär att ju fler kanter en nod har desto större sannolikhet är det att den också får ytterligare nya kanter när nya noder introduceras. Noder med högre grader har således en större förmåga att sammanbindas med nya noder än vad noder med en lägre grad har. Detta gör att när en ny nod introduceras in i grafen har den en preferens för vilka noder den ska binda till, och detta har gett namn åt modellen. Ett adderande av noder enligt denna mekanism leder till att noder med högre grader kommer att få ännu fler kanter och därmed en ännu större grad och därmed en ännu större förmåga att binda till nya noder. Mekanismen leder

också till att skillnader mellan nodernas grader bara ökar med tiden.

Det som gjort modeller baserade på denna mekanism så populära är att de leder till potenslagar för nodgraderna, vilket faktiskt var tanken ända från början. Man tror nämligen att det är just en sådan mekanism i verkliga nätverk som gör att de har en potensfördelning. Man försöker nu efterlika själva uppbyggnaden av nätverket. Det var faktiskt det faktum att så olika nätverk som webben, citationsnätverk och nätverk av skådespelare alla har en potenslagfördelning som fick Barabási och Albert att fundera över en gemensam mekansim som kunde förklara detta, och resultatet blev preferential attachment mekanismen. Och det verkar också trovärdigt, för till exempel ett socialt nätverk, att när en ny individ introduceras in i en gemenskap så är det en större sannolikhet att det uppstår en bekantskap mellan den nya individen och några av de synliga individerna med många bekanta, kanter, än med någon relativt okänd individ. För webben skulle denna preferensmekanism innebära att nyintroducerade internetsidor länkas med en större sannolikhet till mycket välkända internetsidor som till exempel Google än till internetsidor som nästan ingen känner till.

Det finns olika preferential attachment modeller och det är detaljerna i mekanismen som skiljer modellerna åt. En preferential attachment modell som beskrivs i detta kapiel skapas genom algoritmen nedan. Grafen som uppstår efter denna mekanism illustreras i figur 14.

(1) Grafen har från början (t=0) två noder och en kant.
(2) Sedan adderas en nod åt gången vid varje heltalstidpunkt och med den följer en kant. När nod $N+1$ introduceras sammanbinds den andra ändpunkten av kanten till nod i med sannolikheten $d_i(N)/2(N+1)$ där $d_i(N)$ är graden av noden i innan nod $N+1$ adderats. Loopar och multipla kanter är inte tillåtna i grafen.



Figur 14. Utvecklingen av preferential attachment modellen enligt den definierade mekanismen ovan för de tre första nya noder som introduceras in i grafen.

6.1.1 Simulering av preferential attachment modeller

För att simulera en graf där noder introduceras och sammanbinds till andra noder på ett sätt som sker enligt preferential attachment kan man gå tillväga genom att använda sig utav en metod med en lista. Listan visar vilka noder kanterna är fästa vid och måste därför också uppdateras för varje tidssteg, det vill säga efter att varje ny nod introducerats. För varje ny nod väljs slumpmässigt likformigt ur denna lista ett tal för att få fram till vilken nod den nya noden ska binda till.

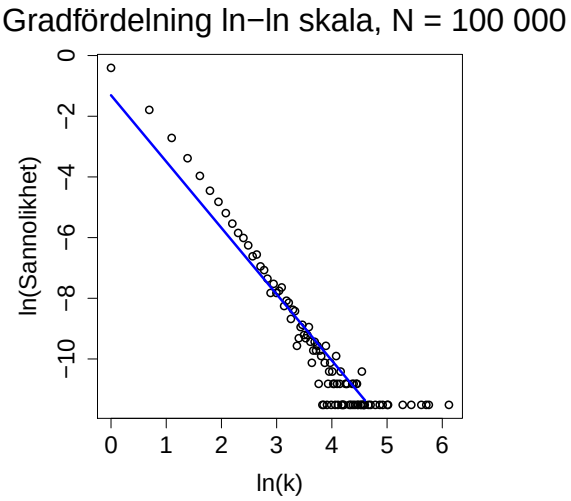
Som ett exempel skulle grafen i figur 14 ovan ha simulerats på följande sätt med denna metod. Först vid t=0 är listan för grafen (1,2). Listan (1,2) visar, genom att nod 1 och nod 2 bara förekommer *en* gång var i listan, att dem båda noderna också bara har *en* grad i grafen, vilket kan ses i figuren. Ur denna lista dras slumpmässigt ett tal, och utfallet blev i detta fall nod 2, vilket innebär att den nya noden ska binda till nod 2. Efter att den nya noden nu sammanbundits och inkluderats in i grafen ges nu listan för grafen vid t=1. Vid t=1 är listan (1,2,2,3) där nod 1 och 3 förekommer en gång var eftersom de båda bara har en grad, och nod 2 förekommer 2 gånger eftersom den har grad 2. Vid t=2 är listan på motsvarnade sätt (1,2,2,3,3,4) och vid t=3 är listan (1,2,2,2,3,3,4,5). Listorna i exemplet är sorterade men det behöver de inte alls vara eftersom talen väljs likformigt slumpmässigt ur listan så att varje tal har en lika stor sannolikhet att väljas ut. Att inte behöva sortera listorna utan låta den byggas på hela tiden blir också enklare praktiskt sätt.

6.2 Preferential attachment modellernas egenskaper

6.2.1 Gradfördelning

Gradfördelningen för preferential attachment modeller följer potenslagfördelningen $p_k \sim k^{-3}$, där exponenten i modellen är $\gamma = 3$ [10]. Eftersom exponenten är oberoende av andra parametrar i modellen kan den inte varieras. Detta är en nackdel för modellen eftersom många nätverk följer en potenslagfördelning med ett något lägre värde än 3 (se tabell 1). Ett nätverk vars gradfördelning dock modelleras med en potenslagfördelning där $\gamma = 3$ är citationsnätverk.

En simulering av preferential attachment modellen definierad med algoritmen beskriven tidigare (se figur 14) simulerades ända tills grafens storlek blev $N = 100\,000$. Simuleringen bekräftar att mekanismen leder till att den realiserade grafen får en gradfördelning som är en tungsvansad potenslagfördelning, se figur 15 nedan.



Figur 15. Simulering av $N = 100\,000$ noder med preferential attachment mekanismen definierad tidigare. Gradfördelningen på logaritmerade axlar följer en rät blå trendlinje, vilket tyder på att gradfördelningen följer en potenslag och trendlinjens lutning anger potensfördelningens exponent, i detta fall $-2.186 = -\gamma$.

6.2.2 Medelnodavstånd

Medelnodavståndet för preferential attachment modeller ökar approximativt logaritmiskt med N, och är [3]

$$l \approx \frac{\ln N}{\ln \ln N} \tag{17}$$

Modellernas medelnodavstånd överensstämmer med de verkliga nätverkens. Vidare har de dessutom ett systematisk kortare medelnodavstånd än vad Erdős-Rényi grafen har.

6.2.3 Klustring

Det finns inget analytiskt resultat för klustringskoefficienten för modellerna. Men eftersom de inte har några triangelskapande mekanismer blir klustringskoefficienten låg och minskar med grafstorleken N. Man kan visa att klustringskoefficienten approximativt följer en potenslag [3]

$$C \sim N^{-0.75} \tag{18}$$

Klustringskoefficienten överensstämmer alltså inte med de verkliga nätverkens.

6.3 Generaliseringar av preferential attachment modeller

Preferential attachment modeller har fått mycket uppmärksamhet vilket lett till att många författare har föreslagit olika tillägg och modifieringar.

En modifiering är när sannolikheten med vilka de nya noderna sammanbinds med de övriga noderna i grafen inte är linjär med graden k av en nod, utan att sambandet istället är exponentiellt, det vill säga k^{α}.

Ytterligare modifieringar man kan göra för att påverka uppbyggnaden av nätverket är att addera kanter, eller också ändra på var kanterna sitter. Man kan också ta bort noder eller kanter.

7 Referenser

[1] Abello J., Pardalos P. M. and Resende M. G. C., (1999), "External memory algorithms and visualization" *DIMACS Series on Discrete Mathematics and Theoretical Computer Science* **50**, 119

[2] Abramowitz M. and Stegun I., *Handbook of Mathematical Functions* (Dover, New York, 1965)

[3] Albert R. and Barabási A.L., (2002), "Statistical mechanics of complex networks" *Reviews of Modern Physics* **74**, 47-97

[4] Albert R., Jeong H. and Barabási A.L., (1999), "Internet: Diameter of the World-Wide Web" *Nature* **401**, 130

[5] Barabási A.L., Jeong H., Ravasz E., Néda Z., Schubert A. and Vicsek T., "Evolution of the social network of scientific collaborations" *Physica A* **311**, 590-614, arXiv:cond-mat/0104162

[6] Deijfen M. and Kets W., (2009), "Random intersection graphs with tunable degree distribution and clustering" *Probability in the Engineering and Informational Sciences* **23**, 661-674

[7] http://www.cs.virginia.edu/oracle/

[8] http://www.orgnet.com/Erdos.html

[9] http://www-personal.umich.edu/ mejn/networks/

[10] Newman M. E. J., "The structure and function of complex networks" *SIAM Review* **45**, 167-256

[11] Newman M. E. J., Strogatz S. H. and Watts D. J., "Random graphs with arbitrary degree distributions and their applications" *Physical Review E*, *SIAM* **64**:026118

[12] Watts D. J. and Strogatz S. H., (1998), "Collective dynamics of small-world networks" *Nature* **393**, 441