Stockholms universitet

Nätverksmodeller med preferential attachment

Bjarne Kampegård

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2010:4
Matematisk statistik
Juni 2010

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2010:4**,
http://www.math.su.se/matstat

Nätverksmodeller med preferential attachment

Bjarne Kampegård*

Juni 2010

Sammanfattning

Det här arbetet behandlar gradfördelningar i nätverk och nätverksmodeller. Ett nätverk kan beskrivas med en graf som innehåller noder och kanter. En nod motsvarar ett element i nätverket och en kant mellan två noder representerar någon typ av koppling mellan motsvarande element. Antalet kanter som en nod har fästa till sig kallas nodens grad. Nya upptäckter hos verkliga nätverk har lett fram till att man börjat överge de gamla nätverksmodellerna och arbeta fram nya. Särskilt har man funnit att gradfördelningarna hos noderna inte blir realistiska i de traditionellt använda modellerna. I detta arbete kommer en ny typ av modell, baserad på så kallad preferential attachement, att behandlas. Det har visat sig att man i denna modell får en asymptotisk gradfördelning som stämmer bra överens med de gradfördelningar som observerats i verkliga nätverk och denna typ av modell har därför övertygat som bra nätverksmodell.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post: bkampegard@hotmail.com. Handledare: Maria Deijfen.

Abstract

This thesis is concerned with degree distributions in networks. Most network models use random graphs as a way to model the uncertainty and the lack of regularity in real networks. A random graph consists of vertices and edges that are placed between the vertices according to some probabilistic rules. When a random graph is used to represent a real network, the vertices represent elements in the network and the edges represent some kind of connection between these elements. The degree of a vertex is the number of edges that are attached to the vertex.

During the last few years, observations on real networks have incited a burst of activity in network modeling. Researchers have found that the old network models do not have realistic properties. In particular, the degree distributions in the models do not agree with the empirical observations. Hence the researchers have begun to search for new network models where the degree distributions become similar to the degree distributions that have been observed in reality.

In this thesis, we will take a look at a special type of model, based on so called preferential attachment. This model type has turned out to offer a convincing explanation for the properties that have been observed in real networks. In particular, its asymptotical degree distribution agrees with what has been observed in real networks.

Förord

Denna uppsats omfattar 15 högskolepoäng och leder till en kandidatexamen i matematisk statistik vid Matematiska Institutionen, Stockholms Universitet.

Jag vill tacka min handledare Maria Deijfen vid Matematiska Institutionen som hjälpt mig med material till arbetet och gett mig tips och goda råd under arbetets gång.

Innehåll

1 INTRODUKTION

Det allmänna intresset för verkliga nätverk har vuxit de senaste åren när bättre datorer har lett fram till att man kan göra utökade studier på området. Vid dessa utökade studier har man kunnat göra observationer som visat på att de modeller som man använt traditionellt för att beskriva nätverk i vissa avseenden inte är realistiska.

Ofta använder man slumpgrafer när man modellerar verkliga nätverk. En graf består allmänt av noder och kanter, och i en slumpgraf sätts kanterna mellan noderna ut efter probabilistiska regler. Antalet kanter som är fästa vid en viss nod kallas för nodens grad och fördelningen för antalet kanter som en nod har i nätverket benämnes nätverkets gradfördelning. I en nätverksmodell motsvarar en nod i grafen ett element i nätverket och en kant mellan två noder representerar en koppling mellan motsvarande element.

Den enklaste formen av slumpgraf är Erdös-Renyi grafen som man får genom att placera en kant mellan varje par av noder oberoende av varandra med en viss sannolikhet p, se tex [3: Kapitel 1]. Denna graf har använts som nätverksmodell traditionellt. Om n betecknar antalet noder i grafen så finns det $n(n-1)/2$ möjliga kanter i grafen, och det förväntade antalet realiserade kanter är $pn(n-1)/2$.

I ett verkligt nätverk antas n vara stort och sannolikheten p liten. Det betyder alltså att det finns ett stort antal element i nätverket och att sannolikheten för att det ska finnas en koppling mellan två slumpvis valda element är liten. Gradfördelningen i grafen skulle då gå mot en Poissonfördelning om man låter n växa mot oändligheten samtidigt som np, den förväntade graden hos en nod, antas vara oförändrad. Om $P(k)$ anger den asymptotiska sannolikheten för att en slumpvis vald nod har grad k och $u = np$ betecknar den förväntade graden skulle man därmed få att

$$P(k) = \exp(-u) \cdot \frac{u^k}{k!}.$$

Detta uttryck för $P(k)$ går mot 0 exponentiellt fort när k blir stort. Observationer hos verkliga nätverk har dock visat att gradfördelningen tenderar att vara mer tungsvansad, det vill säga det finns fler noder med hög grad i verkliga nätverk än vad som föreskrivs av en Poissonfördelning. Gradfördelningen följer ofta en så kallad potenslag, vilket betyder att $P(k)$ är approximativt proportionell mot k^{-r} för någon exponent r.

I detta arbete beskrivs de empiriska observationerna på verkliga nätverk mer noggrannt i kapitel 2. Kapitel 3 handlar om nätverksmodellering, med tonvikt på modeller baserade på så kallad preferential attachement. Detta är

en modelltyp som har visat sig ge upphov till precis den typ av tungsvansade gradfördelningar som har observerats i verkliga nätverk. I kapitel 4 härleds den asymptotiska förväntade gradfördelningen i en preferential attachement modell och det teoretiska resultatet illustreras med hjälp av en simulering.

2 EMPIRISKT OBSERVERADE EGENSKAPER HOS NÄTVERK

Med de bättre och mer utvecklade datorer som har kommit på senare år, så har det blivit möjligt att undersöka uppbyggnaden och strukturen hos verkliga nätverk på ett helt nytt sätt, och det har haft betydelse för nätverksforskningen, se tex [3: Kapitel 1] för en översikt.

Man har bland annat kunnat slå fast att många nätverk är sk "small worlds" vilket betyder att avståndet mellan två noder, definierat som antalet kanter längs den kortaste vägen mellan noderna, är relativt litet. Denna observation var dock inte helt oväntad.

Vad som överraskade mer vid de empiriska undersökningarna, och som kommer att tas upp här, var att många nätverk var skalfria, vilket innebär att gradfördelningen i nätverket är nästan oberoende av nätverkets storlek. Mer precist så sägs ett nätverk vara skalfritt när $P(k)$, andelen av noderna som har ett gradtal lika med k, är nästan proportionell mot k^{-r} för någon exponent $r > 1$ när k är stort. Man säger att $P(k)$ är proportionell mot en potenslag. Det ska alltså finnas en konstant $C > 0$ sådan att det approximativt gäller att

$$P(k) = Ck^{-r}. \tag{1}$$

för stora k. Exponenten r benämnes potenslagexponent. Notera att i en sannolikhetsfördelning enligt högerledet i (1), så är väntevärdet ändligt om $r > 2$ och variansen är ändlig om $r > 3$.

För att undersöka hur väl gradfördelningen i ett empiriskt nätverk stämmer överens med en potenslagfördelning så kan man använda en loglog-plot av gradsekvensen: Om $P(k) = Ck^{-r}$, så gäller att $\log P(k) = \log C - r \log k$, vilket betyder att om man gör en loglog-plot över gradsekvensen, så bör det i plotten bli en i stort sett rät linje med lutning $-r$.

2.1 Exempel på verkliga nätverk

De ungerska matematikerna Reka Albert och Albert-Laszlo Barabasi [1] har sammanställt vad ett flertal forskare kommit fram till vad gäller egenskaper hos verkliga nätverk.

Två exempel på sociala nätverk som har studerats är mängden filmskådespelare där två skådespelare är länkade (har en kant mellan sig) om de har spelat mot varandra, och vetenskapliga artikelförfattare där två författare är länkade om de har skrivit en artikel tillsammans. Båda dessa typer av nätverk har studerats ingående på olika håll, och de har visat sig ha gradfördelningar som kan approximeras väl med potenslagar.

Vid en undersökning som gjordes 2001 av sexuella nätverk var noderna individer i Sverige. En kant mellan två noder fanns om motsvarande personer hade haft en sexuell relation under föregående år. Man fann i undersökningen att gradfördelningarna både för män och för kvinnor följde potenslagar med exponenterna $r(\text{kvinna})=3.5$ och $r(\text{man})=3.3$.

En cell hos en organism kan ses som ett nätverk av kemikalier, där en kant ges av en kemisk reaktion. En kant mellan två noder antas finnas om motsvarande kemikalier båda två finns med samtidigt i någon kemisk reaktion i cellen. Även sådana nätverk har studerats i celler hos ett flertal organismer, och man har funnit gradfördelningar som följer potenslagar. Potenslagexponenterna har haft värden mellan 2.0 och 2.4.

Ett nätverk som har väckt intresse och studerats flitigt på senare år är nätverket World-Wide Web, www. Noderna är i detta fall webbsidor och en kant mellan två noder finns om en av motsvarande webbsidor länkar till den andra. Detta är ett riktat nätverk, och den webbsida som länkar till en annan sida har en utåtgående kant medan den sida som det länkas till har en inkommande kant. Det var Albert-Laszlo Barabasi och Reka Albert som först gjorde en omfattande studie av nätverket World-Wide Web år 1999. De fann då att gradfördelningen både när det gällde inkommande kanter och utgående kanter följde potenslagar med exponenter $r(\text{in})=2.1$ och $r(\text{ut})=2.5$.

Ett annat exempel på ett stort komplext nätverk är Internet. Skillnaden mellan World-Wide Web och Internet är att World-Wide Web är ett virtuellt nätverk medan Internet är ett fysiskt nätverk, som består av routrar och servrar. Man kan studera nätverket Internet på två olika "nivåer". På den ena nivån är noderna routrar inom en domän och kanterna de fysiska kontakter som finns mellan dem, på den andra nivån handlar det om kontakter mellan domäner, en nod motsvarar då en hel domän bestående av hundratals routrar. När man under åren 1995-2000 gjorde ett antal observationer av nätverket Internet på båda nämnda nivåer, så fann man att gradfördelningarna följde potenslagar. På nivån "inom domän" observerade man värden på exponenten r som var i intervallet 2.3-2.5. På den högre nivån, mellan domäner, låg de observerade exponentvärdena omkring 2.2.

Sammanfattningsvis har det konstaterats att den observerade gradfördelningen påfallande ofta stämmer överens med en potenslagfördelning. Man har börjat söka efter en rimlig förklaring till detta fenomen. Kan man finna en övertygande förklaring till fenomenet så har man kanske fått användbar kunskap och bättre förståelse för hur nätverk fungerar i verkligheten.

3 NÄTVERKSMODELLER

Studiet av nätverksmodeller har fått ökad uppmärksamhet på senare år. Det är av intresse att känna till hur verkliga nätverk är uppbyggda då kännedom om denna uppbyggnad kan ge kunskap om viktiga funktioner. När det gäller till exempel ett nätverk av människor, så kan ökad kännedom om nätverkets uppbyggnad leda till att man får bättre förståelse för hur information går fram och hur sjukdomar sprids etc.

Att få fullständig förståelse för hur ett verkligt nätverk är uppbyggt är i de flesta fall helt omöjligt då verkliga nätverk i regel är alldeles för stora och komplexa. Forskningen har istället kommit att fokusera på en mer lokal beskrivning av noderna och kanterna; hur många noder finns det i nätverket, och hur är de i kontakt med varandra? Uppbyggnaden av ett sådant system sker efter probabilistiska regler och man har försökt sig på att göra passande statistiska modeller för denna uppbyggnad. Det har visat sig att de nätverksmodeller som använts traditionellt är otillräckliga för att kunna beskriva verkliga nätverk på ett bra sätt. Särskilt gäller att gradfördelningarna i de traditionella modellerna inte följer potenslagar, vilket man nu har kunnat slå fast att gradfördelningar i verkliga nätverk ofta gör. Man har därför börjat inrikta sig på att finna nya och bättre modeller, som på ett övertygande sätt kan förklara att nätverken ofta har de egenskaper som man nu har kunnat observera att de har. Se [3: Kapitel 1] för mer detaljer om detta.

De flesta komplexa nätverk är stora och växande, till exempel så består nätverket World-Wide Web, www, av ett mycket stort antal webbsidor och nya sidor tillkommer i snabb takt. Det anses därför lämpligt att ha någon typ av dynamisk modell som förklarar hur nätverket växer till och blir som det blir. Man vill att modellerna asymptotiskt, när antalet noder går mot oändligheten, ska ha de egenskaper som man har funnit i verkliga nätverk.

I detta sammanhang är den observerade egenskapen att många nätverk har gradfördelningar som följer en potenslag mycket viktig. Det är en egenskap som har överraskat forskarna, och kan man få en övertygande förklaring till detta fenomen så har man kanske tagit ett viktigt steg i sökandet efter användbar kunskap om verkliga nätverk. En typ av modeller som har övertygat är sådana som är baserade på Preferential Attachment. Preferential Attachment innebär kort sagt att nya noder ständigt ansluter till nätverket och då med högre sannolikhet fäster sina kanter till redan existerande noder i nätverket som har högt gradtal. Detta innebär bland annat att en nod med många kopplingar snabbt får ännu fler, modelltypen kallas därför även för “Rich-get-Richer-Model.“

3.1 Preferential attachment modeller

Dynamiska modeller för komplexa nätverk beskriver hur nätverken växer till och blir som de blir. Denna mängd av modeller har nu kommit att domineras av Preferential Attachment Modeller, PAMs, se [3: Kapitel 8] och [4]. I en PAM tillkommer ständigt nya noder och kanter genom att de nya noderna ankommer en i taget till nätverket och fäster sina kanter en efter en till redan existerande noder i nätverket efter givna probabilistiska regler, varje ny nod antas ansluta med precis m stycken kanter där m är ett fixt heltal. Sannolikheten för en existerande nod att ta emot en av dessa m kanter, *kantsannolikheten*, är större för den nod som redan har många kanter, det vill säga högt gradtal. Preferential attachment anses vara ett rimligt antagande i många nätverk, till exempel bör det vara större sannolikhet att en ny individ i ett socialt nätverk får kontakt med en social individ som redan har många kontakter, och i nätverket World-Wide Web bör sannolikheten vara större att en nytillkommen webbsida börjar länka till en redan känd och populär webbsida, än att sidan börjar länka till en webbsida som inte är så känd.

En PAM beskriver alltså ett växande nätverk där nya noder ständigt tillkommer och fäster sina kanter till de redan existerande noderna och då med större sannolikhet till existerande noder som redan har höga gradtal, den nya noden ansluter enligt så kallad preferential attachment. I verkliga nätverk har man funnit att det är relativt vanligt med noder som har hög grad, betydligt vanligare än vad som föreskrivs av de traditionellt använda nätverksmodellerna. En möjlig förklaring skulle kunna vara att preferential attachment förekommer i någon form. Detta leder som bekant till att noder med ett redan högt gradtal med stor sannolikhet snabbt får ännu högre grad.

En PAM är baserad på en funktion $f(k)$ av nodens gradtal k som anger sannolikheten för att en ny nod fäster en kant till en nod med grad k, på så sätt att kantsannolikheten för noden är proportionell mot $f(k)$, se [2: Kapitel 1]. För en viss PAM finns det alltså en speciell växande funktion $f(k)$, definierad för positiva heltal. Om $f(k) = k + s$ för någon konstant s så har man en linjär PAM, där kantsannolikheten för en existerande nod i nätverket växer linjärt med nodens gradtal vid tillfället. För en sådan PAM visar det sig att man får en asymptotisk gradfördelning som följer en potenslag med en exponent r som beror av modellens parametrar s och m.

De första att studera PAMs i nätverkssammanhang var Albert-Laszlo Barabasi och Reka Albert. I den modell de studerade är $f(k) = k$. Detta är ett specialfall av en linjär PAM, i vilken konstanten s är 0. För en sådan modell kan man visa att den asymptotiska gradfördelningen blir enligt en po-

tenslag med exponent $r = 3$, se tex [3: Kapitel 8]. Andelen noder med grad k blir alltså proportionell mot k^{-3} när k är stort. Detta förklaras närmare i kapitel 4, där en härledning ges av den förväntade asymptotiska gradfördelningen. När man gjort simuleringar av sådana PAMs har man fått grafer där gradfördelningen följer en potenslag med en exponent nära 3, i enlighet med vad teorin säger.

3.2 En linjär preferential attachment modell

I detta arbete ska vi till största delen begränsa oss till fallet då varje ny nod ansluter med precis en kant till nätverket, dvs. $m = 1$. Vi ger nu en mer precis definition av en linjär PAM med $f(k) = k + s$ och $m = 1$ hämtad ur [3: Kapitel 8] och [4].

Först noterar vi att om kantsannolikheten för varje nod ska vara positiv så måste parametern s uppfylla $s > -1$. Nätverket utvecklas på så sätt att vid varje heltalstidpunkt n så anländer en ny nod till nätverket. Den i:te noden som anländer till nätverket betecknar vi med $v(i)$. Låt $D(i, n)$ beteckna gradtalet för $v(i)$ vid tid n, när n noder finns i nätverket. Kantsannoliketen för $v(i)$ vid tid n blir alltså proportionell mot $D(i, n) + s$ ($n = 1, 2, \ldots$ och $i = 1, \ldots, n$).

Låt nu $G(n)$ beteckna nätverket vid tid n och antag att modellen startar från en graf $G(1)$, som består av en enda nod med en kant som är en själv-loop. Att en kant är en själv-loop betyder att dess båda ändar fäster till samma nod. En själv-loop ökar alltså nodens gradtal med 2. Det totala gradantalet vid tid n kommer därmed att vara $2n$ (vilket förklarar nämnarna i uttrycken nedan). Skriv $v(n+1) \to v(i)$ för att beteckna att $v(n+1)$ ansluter till $v(i)$ ($i = 1, \ldots, n+1$). Med $v(n+1) \to v(n+1)$ menas alltså att en själv-loop sker. Nya noder fäster efter följande probabilistiska regel:

$$P(v(n+1) \to v(i)|G(n)) = \begin{cases} \frac{D(i,n)+s}{n(2+s)+(1+s)} & \text{för } i = 1, \ldots, n; \\ \frac{1+s}{n(2+s)+(1+s)} & \text{för } i = n+1. \end{cases}$$

Man kan också välja att inte tillåta själv-loopar, i så fall får man sätta att

$$P(v(n+1) \to v(i)|G(n)) = \begin{cases} \frac{D(i,n)+s}{n(2+s)} & \text{för } i = 1, \ldots, n; \\ 0 & \text{för } i = n+1. \end{cases} \quad (2)$$

Dessutom låter man modellen börja från en graf $G(2)$ som består av två noder som är sammankopplade med två kanter.

En fördel med den senare modellen utan själv-loopar är att man får en sammanhängande graf. En sammanhängande graf är en graf som består av

en enda komponent, och där det alltså går att ta sig från varje nod till varje annan nod. Alla noder hänger därmed ihop genom kanter, man säger att det finns en jättekomponent.

Det visar sig att den asymptotiska gradfördelningen inte påverkas av om man väljer att tillåta själv-loopar eller inte, i bägge fall får man en asymptotisk gradfördelning enligt en potenslag med exponent $3+s$. Modellen kan också definieras för $m \geq 2$, genom att varje ny nod ansluter med m kanter till existerande noder på så sätt att m stycken noder väljs oberoende enligt (2). Man får då en potenslag med exponent $3 + s/m$ (här måste det gälla att $s > -m$ för att kantsannolikheten alltid ska vara positiv). I nästa kapitel härleds den asymptotiska förväntade gradfördelningen för modellen utan själv-loopar med $m = 1$ och $s = 0$.

4 HÄRLEDNING AV GRADFÖRDELNINGEN I EN LINJÄR PREFERENTIAL ATTACHMENT MODELL

I det här kapitlet härleds den förväntade asymptotiska gradfördelningen för den linjära PAM utan själv-loopar och med $m = 1$ som beskrevs i förra kapitlet (kantsannolikheterna definieras i (2)). Vi ska begränsa oss till $s = 0$. Först beskrivs dock gradfördelningen för allmänt $m \geq 1$ och $s > -m$. För ytterligare detaljer, se [3: Kapitel 8].

För att förenkla framställningen är det på sin plats att definiera Gamma-funktionen $t \mapsto L(t)$ f?r $t > 0$ genom

$$L(t) = \int_0^\infty \exp(-x)x^{t-1}dx. \tag{3}$$

Genom partiell integration går det att visa att $L(t+1) = tL(t)$. Vidare gäller att $L(1) = 1$, vilket ger att $L(n) = (n-1)!$ för $n = 1, 2, \ldots$.

Låt nu $P(k, n)$ beteckna andelen av noderna som har ett gradtal k vid tid n (det vill säga när n noder finns i nätverket) i en linjär PAM utan själv-loopar med parametrar s och m och definiera sekvensen $\{p(k), k = 0, 1, 2, \ldots\}$ genom

$$p(k) = \begin{cases} 0 & \text{för } k = 0, \ldots, m-1; \\ \frac{(2+s/m)(L(k+s)L(m+2+s+s/m))}{L(m+s)L(k+3+s+s/m)} & \text{för } k > m-1. \end{cases} \tag{4}$$

Man kan visa att, för alla $k = 0, 1, 2, \ldots$, så konvergerar $P(k, n)$ i sannolikhet mot $p(k)$ då n går mot oändligheten. Vidare gäller att för varje $m = 1, 2,..$ och $s > -m$ så finns en konstant $C = C(m, s) > 0$ sådan att $p(k)$ dividerat med Ck^{-r} konvergerar mot 1, där $r = 3 + s/m > 2$. Detta innebär att "svansen" i den asymptotiska gradfördelningen beter sig som en potenslag när k blir stort. Allt detta formuleras mer precist i nedanstående sats, som är hämtad ur [3: Theorem 8.2].

Sats 4.1 *För en PAM utan själv-loopar och med $m \geq 1$ och $s > -m$, så som definierats i kapitel 3.2, så finns en konstant $C_1 > 0$ sådan att*

$$P\left(\max |P(k, n) - p(k)| > C_1\sqrt{\frac{\log n}{n}}\right) \to 0 \tag{5}$$

n?r $n \to \infty$. Vidare gäller när $k \to \infty$ att

$$p(k) = C_2 k^{-r}(1 + O(1/k)),$$

där $r = 3 + s/m$ *och* $C_2 = C_2(m,s) = \frac{(2+s/m)L(m+2+s+s/m)}{L(m+s)}$.

I fortsättningen ska vi koncentrera oss på fallet då $m = 1$ och $s = 0$. Man kan då enkelt se att

$$p(k) = \frac{2L(k)L(3)}{L(1)L(k+3)}.$$

Från Stirlings formel följer att

$$\frac{L(x+a)}{L(x)} = x^a(1 + O(1/x)) \quad \text{för } a > 0 \text{ när } x \to \infty.$$

Man får alltså att $p(k) \approx 4k^{-3}$ när k är stort.

Satsen bevisas i två steg. I det första steget visar man att den asymptotiska gradfördelningen alltid blir i närheten av den förväntade, och i det andra steget identifierar man denna förväntade gradfördelning. I resten av detta kapitel ska vi nu identifiera den förväntade gradfördelningen då $m = 1$ och $s = 0$ och visa att den ges av $p(k)$. Idén i härledningen är hämtad ur [4]. För ett fullständigt bevis av Sats 4.1, se [3: Kapitel 8.4-5]

Fixera k och antag att sekvensen $E[P(k,n)]$ konvergerar mot något gränsvärde $q(k)$ när n går mot oändligheten. Det har, som nämnts i början av kapitel 2, visat sig att gradfördelningen ofta konvergerar på detta sätt i verkliga nätverk; den empiriska gradfördelningen har i många nätverk visat sig bli nästan oberoende av nätverkets storlek. Sätt $N(k,n)$ som antalet noder som har grad k vid tidpunkten n och kom ihåg att $G(n)$ betecknar grafen vid denna tidpunkt. Vi ska nu se på $E[N(k,n+1) - N(k,n)|G(n)]$, det vill säga den förväntade förändringen av antalet noder med grad k från tid n till tid $n+1$. Vad kan hända som påverkar differensen?

1. Antalet noder som har grad k kan öka med 1 från tid n till tid $n+1$ genom att den nya noden fäster sin kant till en nod med grad $k-1$ och detta sker, enligt (2), med sannolikhet

   $$\frac{(k-1)N(k-1,n)}{2n}.$$

   Om $k = 1$, så ökar också antalet med 1 från tid n till tid $n+1$ genom att den nya noden får grad 1.

2. Antalet noder som har grad k kan minska med 1 från tid n till tid $n+1$ genom att den nya noden fäster sin kant till en nod med grad k och sannolikheten för detta är

   $$\frac{kN(k,n)}{2n}$$

   .

Ur ovanstående får man att

$$E[N(k,n+1)-N(k,n)|G(n)] = \frac{(k-1)N(k-1,n)}{2n} - \frac{kN(k,n)}{2n} + \mathbf{1}_{(k=1)}. \quad (6)$$

Om man nu tar väntevärdet i bägge led, så får man i vänsterledet att

$$E[E[N(k,n+1) - N(k,n)|G(n)]] = E[N(k,n+1)] - E[N(k,n)].$$

Detta följer av formeln $E[E[X|Y]] = E[X]$. I högerledet får man

$$(k-1)E\left[\frac{N(k-1,n)}{2n}\right] - kE\left[\frac{N(k,n)}{2n}\right] + \mathbf{1}_{(k=1)}$$

vilket är detsamma som

$$\frac{(k-1)}{2}E[P(k-1,n)] - \frac{k}{2}E[P(k,n)] + \mathbf{1}_{(k=1)}.$$

Ett antagande var nu att $E[P(k,n)]$ konvergerar mot något gränsvärde $q(k)$ när n går mot oändligheten. Givet att en sådan konvergens sker, så fås för stora n att

$$E[N(k,n+1)] - E[N(k,n)] = (n+1)E[P(k,n+1)] - nE[P(k,n)] \approx q(k)$$

och

$$\frac{(k-1)}{2}E[P(k-1,n)] - \frac{k}{2}E[P(k,n)] + 1(k=1) \approx \frac{(k-1)}{2}q(k-1) - \frac{k}{2}q(k) + \mathbf{1}_{(k=1)}.$$

Då n går mot oändligheten fås alltså från (6) att

$$q(k) = \frac{(k-1)}{2}q(k-1) - \frac{k}{2}q(k) + \mathbf{1}_{(k=1)}$$

vilket leder till rekursionen

$$q(k) = \frac{(k-1)}{(k+2)}q(k-1) + \frac{2}{k+2} \cdot \mathbf{1}_{(k=1)}.$$

Denna löses genom att man sätter $q(0) = 0$ (eftersom varje ny nod ansluter med en kant) vilket ger

$$\begin{aligned} q(1) &= \frac{2}{3} \\ q(2) &= \frac{1}{4} \cdot \frac{2}{3} \\ q(3) &= \frac{2}{5} \cdot \frac{1}{4} \cdot \frac{2}{3} \\ q(4) &= \frac{3}{6} \cdot \frac{2}{5} \cdot \frac{1}{4} \cdot \frac{2}{3} \end{aligned}$$

Med induktion kan man visa att

$$q(k) = \frac{2}{3} \prod_{j=1}^{k-1} \frac{j}{j+3}$$

och med hjälp av gamma-funktionen kan detta skrivas som

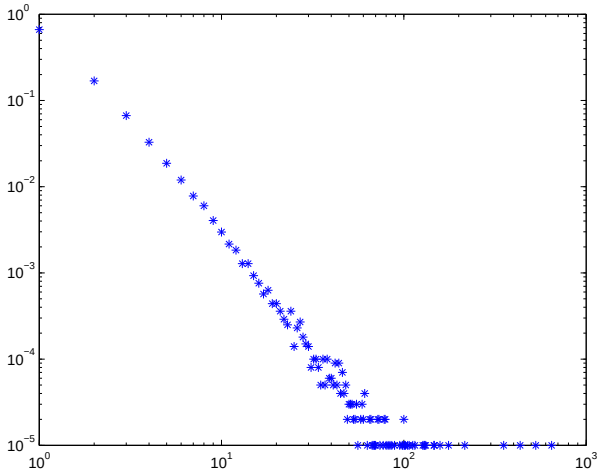
$$q(k) = \frac{2}{3} \frac{L(k)L(4)}{L(k+3)} = \frac{2L(k)L(3)}{L(k+3)} = p(k).$$

Under de antaganden som gjorts, ges alltså den förväntade asymptotiska andelen noder med grad k av $p(k)$, för alla k. Som redan beskrivits avtar $p(k)$ som en potenslag med exponent $r = 3$ då k blir stort.

4.1 Simulering

Figur 1 nedan visar på resultatet vid en simulering enligt den linjära preferential attachment modellen utan själv-loopar, med $m = 1$ och $s = 0$, dvs den modell för vilken den asymptotiska gradfördelningen nu har härletts. Vid simuleringen sattes n till 100 000. Simuleringen startade alltså med två noder som var kopplade till varandra med en kant och nya noder anslöt en i taget efter de givna probabilistiska reglerna (2) ända tills 100 000 noder fanns i grafen.

I figuren är $P(k)$ plottad mot k, och skalorna på x-axeln och y-axeln är logaritmiska, vilket innebär att om gradfördelningen blir enligt en potenslag med exponent 3, så ska det i plotten bli en i stort sett rät linje med lutning -3 (se kapitel 2). Så blir det också när k inte är för stort. För stora k blir det alltför få observationer (dvs få noder med så stor grad) för att få god anslutning till det predikterade värdet. Skalan är också sådan att en liten avvikelse från det predikterade värdet blir mycket tydlig. Det är förklaringen till att punkterna avviker såpass mycket från den givna linjen (med lutning -3) för vissa k när k är stort.



Figur 1: Plott av $P(k)$ mot k på log-log skala för värden på k mellan 1 och 1000. I en viss del av intervallet, när k är under 50-100, lägger sig punkterna approximativt efter en rät linje med en lutning omkring -3. När k blir stort avviker dock punkterna betydligt från denna linje.

Referenser

[1] Albert, R. och Barabasi, A. (2002): Statistical mechanics of complex networks, *Reviews of Modern Physics* 47, s 47-97.

[2] Durett, R. (2006): *Random graph dynamics*, Cambridge University Press.

[3] van der Hofstad, R. (2010): *Random graphs and complex networks*, föreläsningsanteckningar, finns att ladda ner från www.win.tue.nl/ rhofstad.

[4] van der Hofstad, R. (2010): Percolation and random graphs, kapitel 6 i *New Perspectives on Stochastic Geometry* sammanställd av I. Molchanov och W. Kendall.