Normat **52**:3, 122–136 (2004)

Epidemispridning på sociala grafer

Maria Deijfen

Matematisk Statistik
Stockholms Universitet
SE–106 91 Stockholm
mia@math.su.se

1 Inledning

Med epidemispridning menas allmänt spridning av någon slags smitta i någon typ av population. Tänkbara exempel är en infektion som sprider sig bland invånarna i en stad och ett rykte som sprider sig i en skolklass. För att beskriva spridningen matematiskt behöver vi en *modell*, dvs en uppsättning regler som specificerar smittspridningsdynamiken. Beroende på om dessa regler involverar slumpmässighet eller ej sägs modellen vara antingen *stokastisk* eller *deterministisk*. Eftersom det är naturligt att tänka sig att en smittmekanism innefattar ett element av slump – det är sällan vi säkert vet att en individ kommer att bli smittad – är det rimligt att anta att en stokastisk modell bättre fångar beteendet hos en epidemi. Priset för denna högre grad av realism är att en stokastisk modell i regel är betydligt mer komplicerad att analysera än en deterministisk och en rad förenklande antaganden måste införas för att modellen ska kunna hanteras. I majoriteten av de stokastiska epidemimodeller som har studerats antas att populationen i vilken epidemin äger rum är (a) *sluten*, dvs inga dödsfall/födslar och ingen immigration/emigration äger rum; (b) *homogen*, dvs alla individer är av samma typ vad gäller smittsamhet/mottaglighet; (c) *homogent blandad*, dvs en given individ har med samma sannolikhet kontakt med varje annan individ.

I den här uppsatsen ska vi studera den enklaste stokastiska epidemimodellen, *Reed & Frost-modellen*, och försöka släppa på antagandet om homogen blandning i populationen. Att en population är homogent blandad innebär att den saknar social struktur, vilket naturligtvis är mycket orealistiskt. I själva verket består

en mänsklig population av en rad sociala grupperingar – hushåll, arbetsplatser, skolklasser etc – där kontakter sker i långt större utsträckning än i den övriga populationen. För att representera denna sociala struktur ska vi använda grafer, där noderna representerar individer och kanterna representerar sociala relationer. Smittan i fråga sprids sedan längs detta nätverk.

De populationer som betraktas i epidemimodellering är typiskt mycket stora, vilket betyder att det är omöjligt att i detalj kartlägga de sociala relationerna. För att modellera den sociala grafen ska vi därför använda *slumpgrafer*, dvs grafer där kanterna är genererade av någon typ av slumpmekanism. Problemställningen som ska behandlas är tvådelad. Dels vill vi hitta en slumpmekanism som ger grafer som liknar verkliga sociala nätverk, dels vill vi undersöka hur smittspridningen påverkas av den underliggande grafen. De epidemiska storheter vi kommer att intressera oss för är:

- *Reproduktionstalet*, R_0. En kritisk storhet som beror av modellens parametrar och som är definierad så att ett stort utbrott – dvs ett utbrott av samma storleksordning som hela populationen – är möjligt om och endast om $R_0 > 1$. I de modeller vi ska studera ges R_0 av det förväntade antalet nya smittfall som genereras av en given smittad individ i en stor mottaglig population.
- *Epidemins slutstorlek*, τ, som anger hur stor andel av populationen som har upplevt smittan när epidemin upphör.

Resten av uppsatsen är upplagd så att avsnitt 2 innehåller en beskrivning av den vanliga Reed & Frost-modellen med en homogent blandad population, i avsnitt 3 modifieras modellen så att social struktur i form av en underliggande graf inkluderas och i avsnitt 4–6 behandlas ett antal olika slumpgrafkonstruktioner och deras effekter på epidemiprocessen. Läsaren förutsätts vara bekant med grundläggande begrepp inom sannolikhetsteorin. Referenser förekommer relativt sparsamt – för vidare läsning hänvisas till Andersson och Britton (2000) samt, när det gäller epidemispridning på grafer, till Andersson (1999) och Deijfen (2000).

2 Reed & Frost-modellen

En av de första stokastiska epidemimodellerna introducerades 1928 av Reed och Frost. Modellen beskriver spridning av en smitta i en sluten, homogen population som vid tid $t = 0$ består av n osmittade mottagliga individer och en smittsam individ. Dynamiken är följande: Antag att en individ i är smittsam vid tid t ($t = 0, 1, 2, \ldots$). En given mottaglig individ j smittas av i med sannolikhet γ/n, där $\gamma > 0$, och blir, i händelse av en smittöverföring, smittsam vid tid $t + 1$. Individ i avlägsnas från den epidemiska processen vid tid $t + 1$ – tex till följd av immunitet eller isolering – och deltar sedan inte vidare i spridningsförloppet. Alla kontakter antas ske oberoende av varandra och epidemin upphör när det inte finns några smittsamma individer kvar i populationen.

Enligt ovanstående beskrivning förutsätter Reed & Frost-modellen en latent period om en tidsenhet följd av en mycket kort smittsam period, under vilken en given smittsam individ för smittan vidare till en given mottaglig individ med sannolikhet

γ/n. De smittsamma individerna kan alltså delas in i generationer och vi ska nu beskriva hur de inledande faserna i denna generationsprocess kan approximeras av en så kallad *förgreningsprocess* om populationen är stor.

För att förklara dynamiken i en förgreningprocess, betrakta utvecklingen av en ätt där varje individ, under sin livstid, föder ett slumpmässigt antal barn. Dessa barn föder i sin tur ett slumpmässigt antal barn, osv. En förgreningsprocess $\{X_k\}_{k\geq 0}$ anger antalet individer i generation k, dvs det sammanlagda antalet barn som produceras av individerna i generation $k-1$. Reglerna för reproduktionen är att alla individer föder barn enligt samma fördelning, oberoende av varandra. Vanligtvis antas att ätten vid tid $t=0$ består av en enda individ – en stamfader – varifrån alla andra individer härstammar.

I förgreningsprocesstolkningen av en Reed & Frost-epidemi fungerar den individ som är smittsam vid tid $t=0$ som stamfader och en födelse i förgreningsprocessen svarar mot uppkomsten av en ny smittsam individ i epidemiprocessen. Eftersom det från början finns n stycken mottagliga individer som var och en smittas med sannolikhet γ/n, så är antalet nya smittfall som genereras av den initialt smittade individen binomialfördelat med parametrar n och γ/n, dvs sannolikheten att precis k nya smittfall genereras ges av

$$b_k = \binom{n}{k}\Big(1-\frac{\gamma}{n}\Big)^{n-k}\Big(\frac{\gamma}{n}\Big)^k, \qquad k=0,1,\ldots,n.$$

För en given smittsam individ i en senare generation är fördelningen för antalet nya smittfall inte riktigt densamma, eftersom en del av de n initialt mottagliga individerna nu redan har blivit smittade och därmed inte längre deltar i processen. Om vi fortfarande befinner oss i början av epidemiprocessen och om populationen är stor, så är den avlägsnade andelen av populationen dock mycket liten och fördelningen för antalet nya smittfall som genereras av en given smittsam individ approximeras väl av en binomialfördelning med parametrar n och γ/n. Då n är stort kan denna fördelning ersättas av en Poissonfördelning med parameter γ – det gäller ju att

$$b_k \approx \frac{n^k}{k!}\Big(1-\frac{\gamma}{n}\Big)^{n-k}\frac{\gamma^k}{n^k} \longrightarrow \frac{\gamma^k}{k!}e^{-\gamma} \qquad \text{då } n\to\infty,$$

där högerledet känns igen som en Poisson-sannolikhet. Sammanfattningsvis har vi ”visat” att en Reed & Frost-epidemi i en stor population under de inledande faserna beter sig som en förgreningsprocess där avkomman är Poissonfördelad med parameter γ.

I denna uppsats ska vi uteslutande fokusera oss på epidemispridning i stora populationer. Detta betyder att asymptotiska resultat, dvs resultat härledda då $n\to\infty$, kan antas gälla med god noggrannhet. Det förklarar också varför smittsannolikheten, p_s, måste skalas med n: Enligt ovanstående resonemang genererar en smittsam individ i genomsnitt np_s nya smittfall. Om $p_s \equiv p$, gäller förstås att $np_s\to\infty$ då $n\to\infty$, vilket medför att epidemin exploderar och med sannolikhet 1 drabbar hela populationen – en tämligen ointressant modell. Om däremot $p_s=\gamma/n$, är $np_s=\gamma$ oberoende av n och modellen blir icke-trivial även i stora populationer. Låt oss till exempel härleda asymptotiska uttryck för reproduktionstalet, R_0, och slutstorleken. Följande grundläggande sats från förgreningsprocessteorin kommer att vara till hjälp.

Sats 1 *Betrakta en förgreningsprocess där varje individ producerar i genomsnitt γ barn och låt Z beteckna det sammanlagda antalet individer som genereras i processen. Då gäller:*

(i) $\gamma \leq 1 \Rightarrow P(Z < \infty) = 1$;

(ii) $\gamma > 1 \Rightarrow P(Z = \infty) > 0$.

Reproduktionstal: Som beskrivits ovan kan de inledande faserna av den generationsprocess som definieras av de smittsamma individerna i en Reed & Frost-epidemi approximeras av en förgreningsprocess där varje individ föder ett Poissonfördelat antal barn med väntevärde γ. Om $\gamma \leq 1$ är, enligt Sats 1, den totala avkomman i denna förgreningsprocess ändlig med sannolikhet 1. Detta betyder att förgreningsprocessen så småningom dör ut och ett stort utbrott i epidemiprocessen är därmed omöjligt. Om $\gamma > 1$ däremot, finns en positiv sannolikhet att förgreningsprocessen exploderar, vilket ger upphov till ett stort utbrott i epidemiprocessen. Alltså har vi $R_0 = \gamma$ i Reed-Frost modellen.

Slutstorlek: Sannolikheten att en given mottaglig individ undgår att smittas av en given smittsam individ är $1 - \gamma/n$ och för att undslippa epidemin måste den mottagliga individen undgå smitta från samtliga $n\tau$ individer som är smittsamma under epidemins gång. Sannolikheten att en given individ aldrig smittas är alltså $(1-\gamma/n)^{n\tau}$ vilket konvergerar mot $e^{-\gamma\tau}$ då $n \to \infty$. Asymptotiskt ska sannolikheten att en given individ undgår smitta vara lika med andelen av populationen som undgår smitta, och vi får alltså

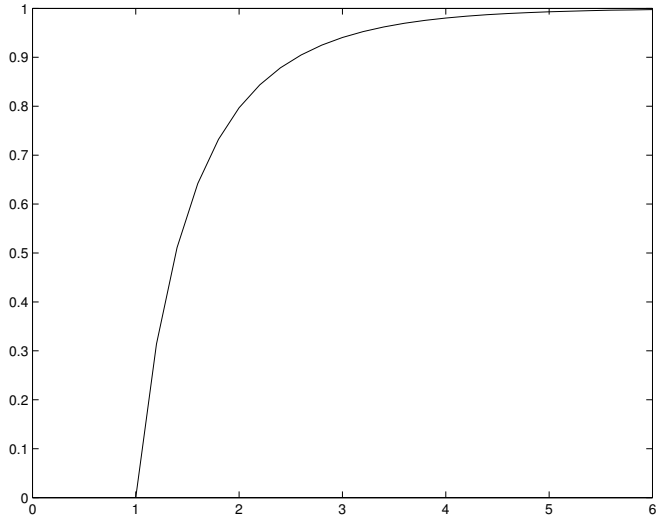
$$1 - \tau = e^{-\gamma\tau}. \tag{1}$$

För $\gamma \leq 1$ är $\tau = 0$ den enda lösningen till denna ekvation. Detta reflekterar det faktum att stora utbrott inte är möjliga då $\gamma \leq 1$. Om $\gamma > 1$ finns en positiv sannolikhet att ett stort utbrott inträffar och det visar sig att ekvationen (1) för slutstorleken i detta fall även har en icke-trivial lösning i intervallet $(0, 1]$. Denna lösning, som ses plottad mot γ i Figur 1, anger slutstorleken i händelse av ett stort utbrott.

3 Reed & Frost-epidemier på grafer

I Reed & Frost-modellen för en given smittsam individ smittan vidare till samtliga andra individer med samma sannolikhet γ/n, vilket i en stor population är ett mycket litet tal. Detta stämmer naturligtvis dåligt överens med dynamiken i en verklig epidemi. En mer realistisk modell vore att låta en smittsam individ smitta de individer hon faktiskt har kontakt med – familj, vänner, arbetskamrater etc – med någon sannolikhet p som inte beror av populationsstorleken. Vi ska nu beskriva hur Reed & Frost-modellen, genom att en social graf införs, kan anpassas så att detta blir fallet.

En graf $\mathcal{G}$ består av en uppsättning *noder*, $\mathcal{V} = (v_1, \ldots, v_n)$, och ett antal *kanter*, $\mathcal{E} = (e_1, \ldots, e_k)$, som sammanbinder par av noder. Varje kant kan skrivas som ett



Figur 1: Slutstorlek vid stort utbrott i en Reed-Frost epidemi, som funktion av γ.

par av noder (v_i, v_j). Två noder tillhör samma *komponent* i grafen om det finns en *stig* som sammanbinder dem, dvs om det går att ta sig från den ena noden till den andra via kanter i grafen. Om den kortaste stigen mellan två noder v_i och v_j består av en enda kant, dvs om $(v_i, v_j) \in \mathcal{E}$, så sägs v_i och v_j vara *grannar*. Slutligen definierar vi *graden* hos en nod som antalet grannar till noden.

En graf kan användas för att representera social struktur i en population och kallas då för ett *sociogram*. I ett sociogram representerar varje nod en individ och kanterna svarar mot sociala relationer mellan individerna. Antag nu att vi vill studera spridningen av en smitta i en sluten, homogen population av storlek n där den sociala strukturen beskrivs av sociogrammet $\mathcal{G}$. Reed & Frost-modellen kan då modifieras på följande sätt: Vid tid $t = 0$ introducerar vi smittan i populationen genom att smitta ner en slumpmässigt vald individ. En individ i som är smittsam vid tid t ($t = 0, 1, 2, \ldots$) smittar sedan ner en given mottaglig granne i $\mathcal{G}$, j, med sannolikhet p, och i händelse av en smittöverföring blir j smittsam vid tid $t + 1$. Individen i avlägsnas från den epidemiska processen vid tid $t + 1$ – tex till följd av immunitet eller isolering – och deltar sedan inte vidare i spridningsförloppet. Alla kontakter antas ske oberoende av varandra och epidemin upphör när det inte finns några smittsamma individer kvar i populationen.

Denna modell påminner mycket om modellen från avsnitt 2, men det finns två viktiga skillnader: En smittsam individ kan bara smitta ner sina grannar i det sociala nätverket och smittsannolikheten är inte skalad med populationsstorleken. Att endast grannar i den sociala grafen kan smittas betyder att populationen inte längre är homogent blandad och modellen kallas därför ibland för den *heterogena Reed & Frost-modellen*. I den homogena Reed & Frost-modellen, där en smittsam individ kan infektera samtliga mottagliga individer i populationen, var vi tvungna att skala smittsannolikheten med populationsstorleken för att förhindra epidemin från att explodera. I denna nya formulering är detta inte nödvändigt, förutsatt att

den sociala grafen har begränsade gradtal, så att antalet grannar till en individ alltså är litet i förhållande till populationsstorleken.

Eftersom vi intresserar oss för epidemier i stora populationer är det omöjligt att i detalj ta reda på hur sociogrammet ser ut. I denna uppsats ska vi låta kanterna som representerar bekantskapsstrukturen genereras av en slumpmekanism och vi vill förstås att den graf vi får ska likna ett verkligt sociogram. Här är några egenskaper vi strävar efter.

– *Begränsad grad*: Det förväntade antalet grannar till en given nod ska förbli ändligt då $n \to \infty$. I ett sociogram reflekterar detta det faktum att bekantskapskretsar är begränsade i storlek även i stora populationer.
– *Transitivitet*: Vi vill att grafen ska innehålla många trianglar. Detta är en av de mest utmärkande egenskaperna hos sociala nätverk och förklaras av att vi kan förvänta oss att många av våra vänner är bekanta också med varandra. Mekanismen som genererar kanterna bör alltså vara sådan att sannolikheten att en viss kant finns med i grafen, givet grafens utseende i övrigt, är större om noderna som den förbinder har en gemensam granne.
– *Realistisk beroendestruktur*: Grafen får inte uppvisa onaturliga beroenden mellan kanterna. Hurvida en viss kant finns med i grafen eller ej bör tex inte påverkas av information om delar av grafen som ligger långt ifrån kanten i fråga. Detta eftersom en individs sociala beteende normalt inte påverkas av individer som hon inte har någon slags anknytning till.

Resten av denna uppsats kommer att ägnas åt exempel på olika typer av slumpgrafkonstruktioner. Vi studerar hur grafen påverkar epidemispridningen i en heterogen Reed & Frost-modell samt undersöker om den kan anses utgöra en god modell för ett socialt nätverk.

4 Bernoulligrafer och triangelgrafer

Den enklaste tänkbara slumpgrafmodellen är *Bernoulligrafen*. Givet $n+1$ stycken noder genererar man en Bernoulligraf genom att, oberoende av varandra, inkludera var och en av de $\binom{n+1}{2}$ möjliga kanterna med sannolikhet r. För en given nod v_i finns n möjliga grannar och var och en av dessa är förbundna med v_i med sannolikhet r. Antalet grannar till en given nod är alltså binomialfördelat med parametrar n och r. För att den förväntade graden, nr, ska förbli ändlig då $n \to \infty$ låter vi $r = \lambda/n$ för något $\lambda > 0$.

Låt oss härleda asymptotiska uttryck för reproduktionstalet och slutstorleken hos en Reed & Frost-epidemi på en Bernoulligraf.

Reproduktionstal: I en stor population är kontaktade individer med stor sannolikhet mottagliga i början av epidemin och de inledande faserna av generationsprocessen av smittsamma individer beter sig därför ungefär som en förgreningsprocess. Enligt Sats 1 finns en positiv sannolikhet att denna förgreningsprocess exploderar (och därmed ger upphov till ett stort utbrott i epidemiprocessen) om och endast om varje individ i genomsnitt föder mer än ett barn. Den kritiska parametern R_0 för

epidemiprocessen fås följaktligen som reproduktionsmedelvärdet i förgreningsprocessen, vilket i epidemitermer motsvaras av det förväntade antalet nya smittfall som genereras av en smittsam individ i början av epidemin.

För att hitta ett uttryck för R_0, betrakta en given smittsam individ i i början av en Reed & Frost-epidemi i en stor population där bekantskapsstrukturen representeras av en Bernoulligraf $\mathcal{G}$. Låt M_i beteckna antalet mottagliga grannar till i i $\mathcal{G}$. Var och en av dessa grannar smittas av i med sannolikhet p, så antalet nya smittfall genererade av i är binomialfördelat med parametrar M_i och p. Det går att visa att väntevärdet i denna fördelning är $\mathrm{E}[M_i]p$. Eftersom populationen är stor är kantsannolikheten i Bernoulligrafen liten och det är därför osannolikt att i har en bekantskapskant till någon av de få individer som inte längre är mottagliga, bortsett förstås från den individ varifrån smittan kom. Antalet mottagliga grannar, M_i, är alltså ungefär binomialfördelat med parametrar $n-1$ och λ/n och eftersom n är stort approximeras denna fördelning väl av en Poissonfördelning med parameter λ. Vi har alltså $\mathrm{E}[M_i] = \lambda$ och således $R_0 = \lambda p$.

Slutstorlek: Låt A_i beteckna händelsen att en given individ i undgår epidemin och låt D_i beteckna antalet grannar till i i $\mathcal{G}$. Vi har då att

$$P(A_i) = \mathrm{E}\big[P(A_i|D_i)\big] = \mathrm{E}\big[P(A_i^j)^{D_i}\big],$$

där A_i^j är händelsen att i undslipper smitta från en given granne j. Asymptotiskt är sannolikheten att j blir smittad densamma som andelen av populationen som drabbas av epidemin, dvs τ, och om j blir smittad så för hon smittan vidare till i med sannolikhet p. Det följer att $P(A_i^j) = 1 - p\tau$ och alltså

$$P(A_i) = \mathrm{E}\big[(1-p\tau)^{D_i}\big]. \tag{2}$$

Den sannolikhetsgenererande funktionen för en stokastisk variabel X definieras allmänt som $\varphi_X(k) = \mathrm{E}[k^X]$ och, om X är Poissonfördelad med parameter α, så fås att $\varphi_X(k) = e^{-\alpha(1-k)}$. Högerledet i (2) känns igen som den sannolikhetsgenererande funktionen för D_i i punkten $1-p\tau$ och, eftersom D_i är asymptotiskt Poissonfördelad med parameter λ, så har vi

$$P(A_i) = e^{-\lambda p\tau}.$$

Asymptotiskt ska sannolikheten att en given individ undgår smitta vara lika med andelen av populationen som slipper undan epidemin. Epidemins slutstorlek, τ, bestäms alltså av ekvationen

$$1 - \tau = e^{-\lambda p\tau}.$$

För givna värden på p och λ löses denna ekvation enkelt numeriskt.

Är nu en Bernoulligraf en bra modell för ett socialt nätverk? Eftersom kantsannolikheten är skalad med populationsstorleken är den asymptotiska genomsnittsgraden i grafen ändlig, men hur är det med de övriga kraven vi ställde upp på s. 127? Låt oss tex undersöka grafens transitivitet – minns att sociala nätverk har en hög grad av transitivitet eftersom sannolikheten att två individer lär känna varandra är stor om de har en gemensam bekant. I en Bernoulligraf är ju dock sannolikheten att två

noder sammanbinds av en kant λ/n oavsett om noderna har en gemensam granne eller ej, eftersom alla kanter inkluderas oberoende av varandra. Detta betyder att grafen asymptotiskt kommer att innehålla mycket få trianglar och alltså inte kan anses vara en god modell för ett socialt nätverk.

Ett försök att öka grafens transitivitet är följande: Förutom att inkludera varje möjlig kant med sannolikhet $r = \lambda/n$, inkluderar vi, oberoende av varandra, även var och en av grafens $\binom{n+1}{3}$ möjliga trianglar med någon sannolikhet $\tilde{r}$. Detta ger en graf $\mathcal{G}$ som kan skrivas som unionen av en Bernoulligraf $\mathcal{G}_1$ och en triangelgraf $\mathcal{G}_2$. Multipla kanter mellan två noder i denna konstruktion reduceras till en, så att varje nodpar alltså förbinds av högst en kant.

Kantsannolikheten i Bernoulligrafen är skalad med populationsstorleken så att $\mathcal{G}_1$ har genomsnittsgrad λ för alla n. För att den asymptotiska genomsnittsgraden i $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$ ska vara ändlig krävs att även triangelsannolikheten $\tilde{r}$ skalas på rätt sätt. Graden för en given nod v i $\mathcal{G}_2$ är $2X$, där X betecknar antalet trianglar som har ett av sina hörn i v. Eftersom de andra två hörnen ska väljas bland övriga n noder finns $\binom{n}{2}$ möjliga trianglar som innehåller v och var och en av dessa trianglar inkluderas i $\mathcal{G}_2$ med sannolikhet $\tilde{r}$. Alltså är X binomialfördelad med parametrar $\binom{n}{2}$ och $\tilde{r}$, och för att väntevärdet i denna fördelning ska förbli ändligt då $n \to \infty$ låter vi $\tilde{r} = \tilde{\lambda}/\binom{n}{2}$ för något $\tilde{\lambda} > 0$.

I Deijfen (2000) härleds asymptotiska uttryck för reproduktionstalet, R_0, och slutstorleken, τ, för en Reed & Frost-modell på den graf som beskrivs ovan. Det visar sig att

$$R_0 = \lambda p + 2\tilde{\lambda}(p + p^2 - p^3)$$

och för slutstorleken fås ekvationen

$$1 - \tau = \exp\big\{-\lambda\tau p - \tilde{\lambda}\big(2\tau(2-\tau)p + \tau(2-3\tau)p^2 - 2\tau(1-\tau)p^3\big)\big\}.$$

För detaljer hänvisas till Deijfen (2000).

Via parametern $\tilde{\lambda}$ kan vi styra andelen trianglar i en triangelgraf och bristen på transitivitet hos Bernoulligrafen kan alltså kompenseras genom att $\tilde{\lambda}$ väljs tillräckligt stor. Det visar sig dock att triangelgrafen har en beroendestruktur som inte är riktigt naturlig i ett socialt nätverk. För att förstå detta, betrakta sannolikheten att det finns en kant mellan två givna noder v_i och v_j i en triangelgraf, givet resten av grafen (dvs hela konfigurationen av kanter, undantaget kanten mellan v_i och v_j, är given). Om kanten (v_i, v_j) inte skapar en triangel är sannolikheten att den finns med i grafen 0. Antag å andra sidan att kanten faktiskt skapar en triangel, dvs att v_i och v_j har en gemensam granne v_k. Om kanterna (v_i, v_k) och (v_j, v_k) redan ingår i andra trianglar så är den betingade sannolikheten att kanten (v_i, v_j) finns med i grafen

$$1 - \left(1 - \tilde{\lambda} \Big/ \binom{n}{2}\right)^{n-1} \approx 2\tilde{\lambda}/n,$$

eftersom minst en av de $n-1$ möjliga trianglarna med hörn i v_i och v_j måste inkluderas. Ingår (v_i, v_k) och (v_j, v_k) däremot inte i andra trianglar så finns det med sannolikhet 1 en kant mellan v_i och v_j. Sannolikheten att två individer i och j är bekanta med varandra påverkas alltså inte bara av om de har en gemensam bekant eller ej (dvs av om en kant mellan v_i och v_j skulle skapa en triangel eller ej),

utan även av bekantskapsformationer som inte är gemensamma för de två individerna (ingår v_i och v_j i andra trianglar eller ej?). Sådan information är normalt inte relevant i en social graf och, även om fenomenet mildras något av Bernoulligrafen, så minskar detta beteende lämpligheten hos denna sammansmältning av en Bernoulligraf och en triangelgraf som modell för ett socialt nätverk.

5 Markovgrafer

Markovgrafen är ett försök att skapa en transitiv grafmodell utan den typ av onaturliga beroenden mellan kanterna som triangelgrafen uppvisar. Definitionen av modellen ges i form av ett sannolikhetsmått som innehåller en parameter som styr antalet trianglar i grafen och där kanter som inte har någon gemensam nod är oberoende. Tyvärr visar det sig att modellen har ett mycket onaturligt asymptotiskt beteende. I detta avsnitt är det viktigt att skilja mellan en slumpgraf som stokastisk modell för generering av kanter och en konkret realisering av denna. Vi kommer därför att reservera beteckningen $\mathcal{G}$ för den stokastiska modellen och skriva G för realiseringar.

En graf med n noder kan beskrivas av *kantindikatorer* $\{I_{ij}\}$ $(i, j = 1, \dots n)$ definierade så att

$$I_{ij} = \begin{cases} 1 & \text{om } (v_i, v_j) \in \mathcal{E} \\ 0 & \text{annars,} \end{cases}$$

dvs $I_{ij} = 1$ om det finns en kant mellan noderna v_i och v_j och $I_{ij} = 0$ annars. I en slumpgrafmodell $\mathcal{G}$ är kantindikatorerna stokastiska variabler och vi kan införa en *beroendegraf*, $\mathcal{D}$. Detta är en icke-stokastisk graf som konstrueras genom att varje kantindikator representeras med en nod och en kant ritas mellan två noder om motsvarande indikatorer är beroende givet övriga indikatorer. Noderna i $\mathcal{D}$ är alltså de möjliga kanterna i $\mathcal{G}$ och kanterna i $\mathcal{D}$ svarar mot de par av kanter i $\mathcal{G}$ som är betingat beroende.

Med en *klick* menas inom grafteorin en fullständig delgraf, dvs en delmängd av noderna sådan att samtliga möjliga kanter finns med. Klickbegreppet är centralt i följande sats, som specificerar sannolikhetsfunktionen för en slumpgrafmodell $\mathcal{G}$ med given beroendestruktur $\mathcal{D}$.

Sats 2 *Betrakta en given realisering G av en slumpgraf $\mathcal{G}$ med beroendestruktur $\mathcal{D}$ och låt E beteckna kantmängen i G. Grafen G har sannolikhet*

$$P(G) = z^{-1} \exp\Big\{ \sum_{K \subseteq E} \alpha(K) \Big\},$$

där $\alpha(K)$ är en godtycklig konstant om K är en klick i $\mathcal{D}$ och $\alpha(K) = 0$ annars.

De enda egenskaper vi kan styra hos en slumpgraf är alltså de som på något sätt är kopplade till klickar i beroendegrafen. En sådan egenskap kan belönas (bestraffas) genom att konstanterna $\alpha(K)$ väljs så att grafer som uppvisar egenskapen i fråga ges större (mindre) sannolikhet än grafer som inte gör det.

En slumpgrafmodell $\mathcal{G}$ sägs vara *Markovsk* om dess beroendegraf inte innehåller några kanter mellan indikatorer I_{ij} och I_{kl} som svarar mot disjunkta nodpar (kanter). Två kanter i en Markovgraf tillåts alltså vara beroende endast om de har en gemensam nod. Klickarna i beroendegrafen svarar således mot kantmängder där samtliga kanter har en nod gemensam och de enda konfigurationer där detta är fallet är trianglar, $T_{v_iv_jv_k} = \{(v_i, v_j),\ (v_j, v_k),\ (v_k, v_i)\}$, och stjärnor, $S_{v_{i_0}\dots v_{i_k}} = \{(v_{i_0}, v_{i_l});\ l = 1, \dots, k\}$, där $k = 1, \dots n-1$.

Låt oss nu formulera Sats 2 i fallet då $\mathcal{G}$ är Markovsk. Introducera först notationen $\alpha(T_{v_iv_jv_k}) = \xi_{ijk}$ och $\alpha(S_{v_{i_0}\dots v_{i_k}}) = \sigma_{i_0\dots i_k}$. För att förenkla modellen och reducera antalet parametrar antar vi att isomorfa grafer har samma sannolikhet. Detta betyder att vi, istället för att introducera en parameter för varje möjlig triangel, nöjer oss med en enda triangelparameter som styr det totala antalet trianglar i grafen och, på samma sätt, för varje k definierar vi en enda parameter som kontrollerar det totala antalet k-stjärnor. Vi har alltså $\xi_{ijk} = \xi$ och $\sigma_{i_0\dots i_k} = \sigma_k$. Sannolikhetsfunktionen för en Markovgraf fås nu som ett korrollarium till Sats 2.

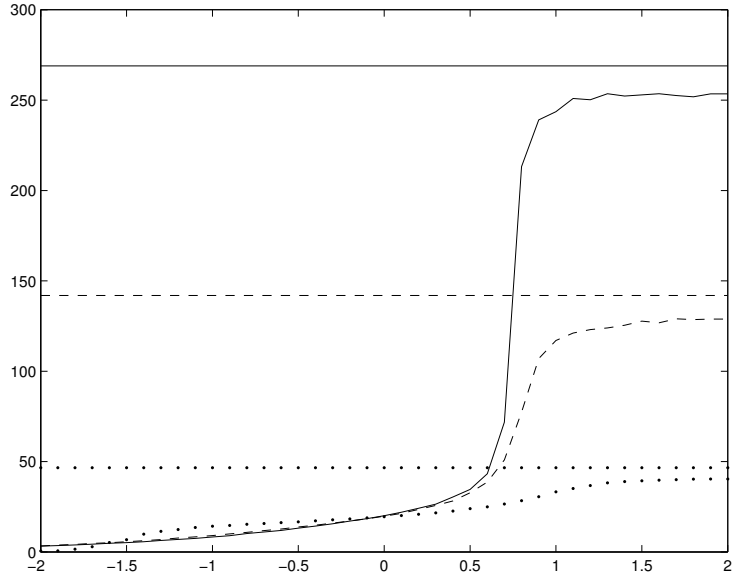
Korrollarium 3 *En given realisering G av en Markovgraf med n noder har sannolikhet*

$$P(G) = z^{-1} \exp\left\{\xi t(G) + \sum_{k=1}^{n-1} \sigma_k s_k(G)\right\},$$

där $t(G)$ är antalet trianglar och $s_k(G)$ antalet k-stjärnor i G.

I en Markovgrafmodell tillåts kanter vara beroende endast om de har en nod gemensam. I termer av bekantskapsformationer betyder detta att en individs sociala beteende endast kan påverkas av individer som hon har någon typ av relation till, vilket ju är en naturlig egenskap hos ett socialt nätverk. Markovgrafmodellen innehåller dessutom en parameter, ξ, som gör det möjligt att styra antalet trianglar i grafen – genom att välja $\xi > 0$ åstadkoms ett mått som belönar grafer med många trianglar. Så långt tycks Markovgrafen vara en mycket lovande modell för ett socialt nätverk. Som nämndes i detta avsnitts inledning visar det sig dock att grafen har ett ofördelaktigt asymptotiskt beteende: I Strauss (1986) visas att, om $\xi > 0$, så gäller att sannolikheten att en godtyckligt stor andel av kanterna i grafen samlas i en enda stor klick konvergerar mot 1 då $n \to \infty$. För ett socialt nätverk betyder detta att samtliga bekantskapskanter i en stor population med stor sannolikhet kommer att vara samlade i en enda jättelik bekantskapskomponent där alla individer är bekanta med varandra. Detta gör Markovgrafen mycket olämplig som nätverksmodell.

Låt oss illustrera Strauss' resultat med hjälp av en simulering. För att förenkla simuleringen har vi valt $\sigma_k = 0$ för alla k och måttet blir alltså $P(G) = z^{-1} \exp\{\xi t(G)\}$. Vidare har vi fixerat grafens genomsnittsgrad. Eftersom det i en graf med n noder och genomsnittsgrad d finns ungefär $nd/2$ kanter, så är detta i princip detsamma som att fixera antalet kanter i grafen. Markovgrafer med n noder och $nd/2$ kanter har genererats med hjälp av MCMC-metoder och det förväntade antalet trianglar i grafen har beräknats genom att utnyttja att tidsmedelvärdet i en Markovkedja konvergerar mot väntevärdet – se Deijfen (2000) för detaljer angående detta. Simuleringarna är mycket tidskrävande varför endast relativt små grafer har studerats – den största grafen har $n = 30$ noder.



Figur 2: Förväntat och maximalt antal trianglar som funktion av ξ i en Markovgraf med 10, 20 respektive 30 noder och genomsnittsgrad 5.

Det maximala antalet trianglar i en graf med ett fixt antal kanter fås förstås om kanterna är samlade i en enda klick. Storleken, $k_{\max}$, hos den maximala klicken fås ur sambandet $\binom{k_{\max}}{2} = nd/2$. Approximativt har vi $k_{\max} = \sqrt{nd}$ och det maximala antalet trianglar i grafen är alltså $\binom{\sqrt{nd}}{3}$. Markovgrafens degenererade beteende illustreras i Figur 2. Vi ser där att, då ξ blir positiv, så närmar sig det förväntade antalet trianglar i grafen snabbt det maximala antalet (som är markerat med räta linjer), vilket betyder att samtliga kanter är samlade i en klick. Omslagspunkten, som för den största grafen med $n = 30$ noder infaller i närheten av $\xi = 0.5$, kryper för större grafer in mot $\xi = 0$.

6 "Small-world"-grafer

Att möta en fullständig främling och upptäcka att man har en gemensam bekant är något som många av oss har varit med om. "Världen är liten!" kanske vi utbrister. Faktum är att detta är ett fenomen som har observerats i många verkliga sociala nätverk: Väljer vi ut två godtyckliga individer så kan de ofta förbindas med en förvånansvärt kort bekantskapskedja, se tex Milgram (1967). Detta är anmärkningsvärt eftersom sociala nätverk för det mesta är (a) stora; (b) glesa, dvs genomsnittsgraden är liten jämfört med populationsstorleken; (c) decentraliserade, dvs de innehåller ingen centralnod till vilken de flesta andra noder är länkade.

"Small-world"-nätverken introducerades av Strogatz och Watts (1998) och är ett försök att skapa transitiva grafer med korta stigar. Transitivitet och korta stigar är egenskaper som i förstone kan tyckas svåra att förena hos en graf, eftersom

transitivitet uppstår då vi ”klumpar ihop” kanterna och korta stigar då vi sprider ut dem, men i det här avsnittet ska vi beskriva en konstruktion som faktiskt lyckas med detta förutsatt att dess parametrar väljs på rätt sätt. För att kvantifiera transitiviteten och stiglängden inför vi följande storheter:

– *Klustringsindex*, C. Antag att en given nod v_i har k grannar. Om $k \geq 2$ finns $\binom{k}{2}$ möjliga kanter mellan dessa grannar. Låt E_{v_i} beteckna antalet av dessa möjliga kanter som finns med i grafen och definera klustringen för noden v_i som

$$C_{v_i} = \begin{cases} \dfrac{E_{v_i}}{\binom{k}{2}} & \text{om } k \geq 2, \\ \dfrac{|\mathcal{E}| - k}{\binom{n}{2} - (n-1)} & \text{om } k < 2. \end{cases}$$

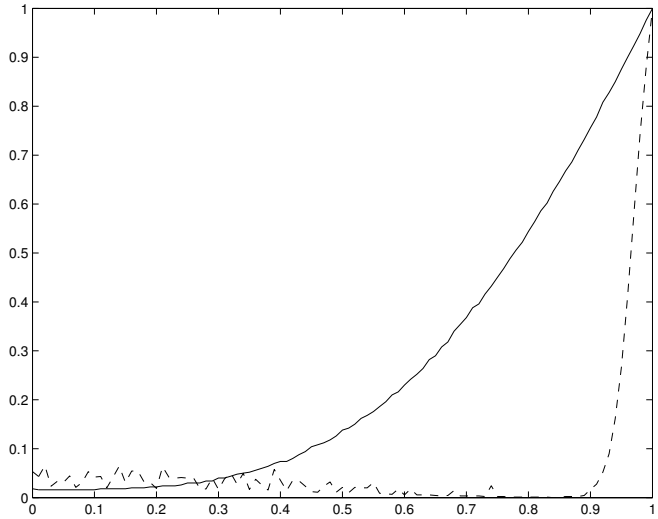
(Definitionen av C_{v_i} för $k < 2$, som vid första anblicken kan verka lite krånglig, är nödvändig för att komma runt problemet med noll-division. Den kan tolkas som den övergripande klustringen i resten av grafen och anger hur stor andel av de totalt $\binom{n}{2} - (n-1)$ möjliga kanterna utan anknytning till noden v_i som är inkluderade i grafen.) Klustringsindexet för grafen definieras nu som

$$C = \frac{1}{n} \sum_{i=1}^{n} C_{v_i}.$$

I ett socialt nätverk är klustringsindexet ett mått på i vilken utsträckning våra vänner är vänner också med varandra.

– *Stiglängdsindex*, L. För en given graf med n noder, låt S_m beteckna antalet nodpar mellan vilka den kortaste stigen består av mer än m kanter ($m = 1, \ldots, n-1$) eller som inte har någon stig alls emellan sig och definiera $L_m = S_m / \binom{n}{2}$, dvs L_m är andelen nodpar i grafen som ligger långt ifrån varandra i den meningen att det krävs mer än m länkar för att ta sig från den ena noden till den andra, om det överhuvudtaget går. Vill vi kunna använda detta mått för att jämföra grafer med olika antal noder bör m väljas som en växande funktion av n: Bekantskapskedjorna är troligen längre i en storstad med 1 000 000 invånare än i en ort med 1 000 invånare. Vi väljer $m = \lfloor \log n \rfloor$ och skriver $L_{\lfloor \log n \rfloor} = L$.

En graf sägs vara ett ”small-world”-nätverk om klustringsindex är stort och stiglängdsindex litet. Vad menas då med ”stort”/”litet” här? Det största möjliga värdet för C är 1, vilket antas i en fullständig graf (dvs en graf där samtliga kanter finns med), och det minsta tänkbara värdet är 0, vilket antas i en tom graf (dvs en graf som saknar kanter). Dessa två grafer är också extrema i fråga om stiglängd. Det är dock inte särskilt instruktivt att jämföra dessa grafer med varandra, eftersom klustringen förstås ökar och stiglängden minskar då fler kanter adderas till en graf. Vi ska här studera grafer med ett fixt antal kanter och se hur klustringen och stiglängden påverkas då vi, genom att justera modellens parametrar, förändrar fördelningen av kanterna över grafen. Målet är att hitta en struktur som ger ett stort värde på C och ett litet värde på L, där ”stort”/”litet” ska ses i relation till det största/minsta möjliga värdet för en graf med det givna antalet kanter.



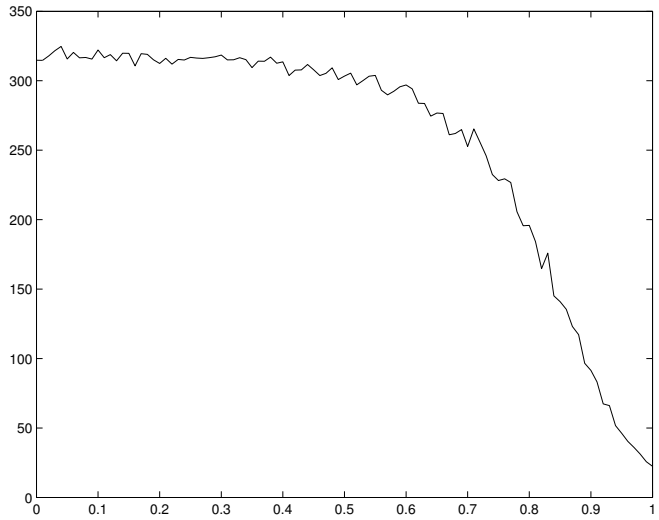
Figur 3: Klustringsindex (heldragen linje) och stiglängdsindex (streckad linje), skalade med värdet för $r = 1$, som funktion av r.

För att konstruera nätverket, arrangera de n noderna i en cirkel. Låt $\mathcal{G}_1$ vara en graf där varje nod sammanbinds med var och en av sina k närmaste grannar åt båda håll med sannolikhet r, där $k \ll n$, och låt $\mathcal{G}_2$ vara en Bernoulligraf med kantsannolikhet λ/n. Grafen $\mathcal{G}_1$ är här tänkt att representera en lokal bekantskapsstruktur (tex det nätverk som uppstår då vi lär känna folk i vårt eget kvarter) och Bernoulligrafen $\mathcal{G}_2$ representerar globala kontakter (tex människor vi lär känna när vi är ute och reser). Den graf vi ska studera är $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$.

Låt d beteckna den asymptotiska genomsnittsgraden i $\mathcal{G}$. Eftersom $k \ll n$ är det, om n är stort, mycket osannolikt att en Bernoullikant ska överlappa en kant i $\mathcal{G}_1$. Genomsnittsgraden i $\mathcal{G}$ ges därför asymptotiskt av summan av genomsnittsgraderna i $\mathcal{G}_1$ och $\mathcal{G}_2$, dvs $d = 2kr + \lambda$. För att reducera antalet fria parametrar i modellen, fixerar vi genomsnittsgraden till $2k$ så att

$$2kr + \lambda = 2k. \tag{3}$$

(Eftersom vi fritt kan välja k när vi konstruerar grafen är den enda inskränkningen här att vi antar att genomsnittsgraden är ett jämnt tal.) Detta är i princip detsamma som att fixera antalet kanter i grafen till nk. Är det nu möjligt att välja parametrarna r och λ så att klustringsindex, C, är stort och stiglängdsindex, L, litet i relation till de extrema värdena för en graf med nk kanter? I Figur 3 visas simulerade klustrings- och stiglängdsindex för olika värden på den lokala kontaktsannolikheten r. Den globala kontaktparametern λ fås ur (3). De simulerade graferna har $n = 500$ noder och genomsnittsgrad $d = 4$ (dvs $k = 2$). Då r minskas från 1 ser vi först ett kraftigt fall för L, följt av ett intervall där L är litet samtidigt som C fortfarande är stort. För r i detta intervall, säg $r \in (0.8, 0.9)$, uppvisar grafen ett ”small-world”-beteende. Minskas r ytterligare avtar C mot sitt låga värde för den rena Bernoulligrafen (som vi har då $r = 0$) och ”small-world”-egenskapen



Figur 4: Genomsnittlig storlek hos den största komponenten i smittgrafen som funktion av mot r då $n = 500$, $k = 2$ och $p = 0.4$.

förstörs. Denna simulering ger vid handen att vi, genom att introducera endast en mycket liten andel globala kanter i nätverket, minskar stiglängden dramatiskt utan att påverka klustringen nämnvärt, med ett ”small-world”-nätverk som resultat.

Reproduktionstal: För $k = 1$ visas i Deijfen (2000) att

$$R_0 = \left(\frac{2pr}{1 - pr} + 1 \right) \lambda p.$$

För att skapa ett realistiskt nätverk krävs dock större värden på k, säg $k \geq 10$, och för sådana k saknas explicita uttryck för R_0 – se dock Moore och Newman (2000:1) för relaterade resultat.

Slutstorlek: Slutstorleken hos en Reed & Frost-epidemi på ett ”small-world”-nätverk har bla studerats av Moore och Newman (2000:2). De analytiska resultaten är dessvärre inte särskilt explicita och här nöjer vi oss med att presentera ett simuleringsresultat.

Givet en social graf och en smittsannolikhet p kan vi tunna ut grafen med sannolikhet $1 - p$, så att varje kant alltså lämnas kvar i grafen med sannolikhet p och suddas ut med sannolikhet $1 - p$. Vi får då en *smittgraf* som representerar utfallet av en Reed & Frost-epidemi: Epidemin kommer att drabba precis de individer som tillhör samma komponent som den initialt smittade individen. I Figur 4 har den genomsnittliga storleken hos den största komponenten i en smittgraf baserad på ett ”small-world”-nätverk plottats mot den lokala kontaktsannolikheten r (den globala parametern λ bestäms av (3)). Detta ska tolkas som slutstorleken för den värsta tänkbara epidemin initierad av en enda individ. Vi ser att komponenterna blir större då r avtar – dvs då andelen globala länkar ökar – och den största förändringen sker just i intervallet $r \in (0.8, 0.9)$ där ”small-world”-beteendet hos grafen sätter in.

7 Avslutande kommentar

Avsikten med denna uppsats har varit att ge en inblick i stokastisk epidemimodellering och att beskriva hur slumpgrafer kan användas för att modellera social struktur. Samtliga grafmodeller vi har tagit upp är behäftade med olika typer av nackdelar: Bernoulligrafen har alltför låg transitivitet, triangelgrafen har en något orealistisk beroendestruktur, i Markovgrafen fördelas kanterna på ett mycket onaturligt sätt och ”small-world”-nätverken är analytiskt svårhanterliga. Sökandet efter goda modeller för sociala nätverk fortsätter!

Bibliografi

Andersson, H. (1999): Epidemic models and social networks, *The mathematical scientist* **24**, 128–147.

Andersson, H. och Britton, T. (2000): *Stochastic Epidemic models and their statistical analysis*, Springer.

Deijfen M. (2000): Epidemics on social network graphs, Examensarbete 2000:1, Stockholms Universitet.

Milgram, S. (1967): The small world problem, *Psychology Today* **2**, 60–67.

Moore, C. och Newman, M. (2000:1): Epidemics and percolation in small-world networks, *Phys. Rev. E* **61**, 5678–5682.

Moore, C. och Newman, M. (2000:2): Exact solution of site and bond percolation on small-world networks, *Phys. Rev. E* **62**, 7059–7064.

Strauss, D. (1985): On a general class of models for interaction, *SIAM Review* **28**, 513–527.

Strogatz, S. och Watts, D. (1998): Collective dynamics of small-world networks, *Letters to Nature* **393**, 440–442.