Introdução a Redes Complexas

Jornadas de Atualização em Informática (JAI)
CSBC 2011

Encontro 2/3

Daniel R. Figueiredo
LAND – PESC/COPPE/UFRJ

Do que trata Redes Complexas?

- Entender como as coisas se conectam e as consequências desta conectividade

"Coisas que se conectam" → **Redes**

"Como, por que, e consequências" → **Complexo**

- ***Estrutura*** assume papel central
  - o necessária para compreender fenômenos que ainda não explicamos
- A rede não é complexa!

Organização do Mini-curso

- Três encontros (Qui 17h, Sex 11h, Sex 17h)
  - duas horas cada, brinde na Sex 17h
- **Encontro 1**
  - Redes por todos os lados, formalismo e propriedades, descobrimentos empíricos
- **Encontro 2**
  - Lei de potência, modelos de redes: G(n,p), Barabási-Albert, Watts-Strogatz, propriedade dos modelos
- **Encontro 3**
  - Funcionalidade em redes: robustez, busca e navigabilidade

Interatividade com participação do público!

Abaixo a Democracia

- Graus dos vértices é muito desigual
- Muitos com grau pequeno, poucos muito grande
- Ex. média 10, mas valores 100, 1000 e 1000 fazem parte da população

Modelo matemático que captura esta característica?

- Distribuição de lei de potência
- Probabilidade não desprezível de ter um elemento muito diferente da média

Lei de Potência

- Qualquer função que tenha a forma

$$f(x) \sim c\,x^{a}$$

c e a são
constantes qq > 0

- f(x) evolui de acordo com uma potência
  - lei de potência! (ou polinômio com grau real)
- Relação ocorre quando x grande

$$f(x) \sim g(x) \longrightarrow \frac{f(x)}{g(x)} \to 1 \quad \text{quando} \quad x \to \infty$$

Distribuição de Lei de Potência

- Lei de potência aplicada a distribuição de probabilidade

$$f_X(x) \sim c\, x^{-a}$$

expoente negativo, a > 0

- onde X é uma v.a. e $f_X(x)$ sua função de probabilidade
- X pode ser discreta ou contínua, limitada ou ilimitada superiormente
- *Cauda pesada*
  - valores grandes podem ocorrer com prob. não desprezível

Distribuição Zeta

- Seja X uma v.a. discreta ilimitada superiormente
- X assume valores 1, 2, ...

$$p_k = P[X = k] \sim k^{-a}$$

- Determinar constante c de normalização

$$p_k = c\,k^{-a}$$

- Temos então

$$\sum_{k>0} p_k = 1 \longrightarrow c = 1 / \sum_{k>0} k^{-a} = 1/z(a)$$

Função Zeta de Riemann

- z(a) é a função zeta de Riemann

$$z(a) = \sum_{k>0} k^{-a}$$

- Problema: determinar quando z(a) converge
- Se z(a) não converge, p_k não é funçao de probabilidade (não temos como normalizar)
- Converge para qualquer a > 0 ?

Converge apenas para a > 1

Momentos da Zeta

- Calcular momentos da distribuição zeta
- Cola: n-ésimo momento é dado por $E[X^n]$
- Primeiro momento é valor esperado

Idéias???

- Aplicar definição

$$E[X^n] = \sum_{k>0} k^n p_k$$

Momentos da Zeta

- n-ésimo momento é dado por

$$E[X^n] = z(a-n)/z(a)$$

Converge sempre?

- para a – n > 1 ou seja, a > n + 1
- infinito caso contrário (não converge)

- Exemplo: a = 2.3, E[X] existe? Var[X] existe?
- Cola: $Var[X] = E[X^2] - E[X]^2$
- Muitas redes tem expoente 2 < a < 3

Invariância em Escala

- Propriedade “scale free”
- Multiplicar o argumento por k mantém a forma de f

$$f(bk) \sim c(bk)^{-a} \sim b^{-a} ck^{-a} = b^{-a} f(k)$$

constante depende apenas de k

- A forma é preservada independente da escala
- Ex. a = -2, k = 3
  - evento 3 vezes maior tem 9 vezes menos probabilidade de ocorrer, independente do tamanho do evento!

Escala log-log

- Aplicar logaritmo aos dois lados da equação

$$f(k) \sim c\, k^{-a}$$

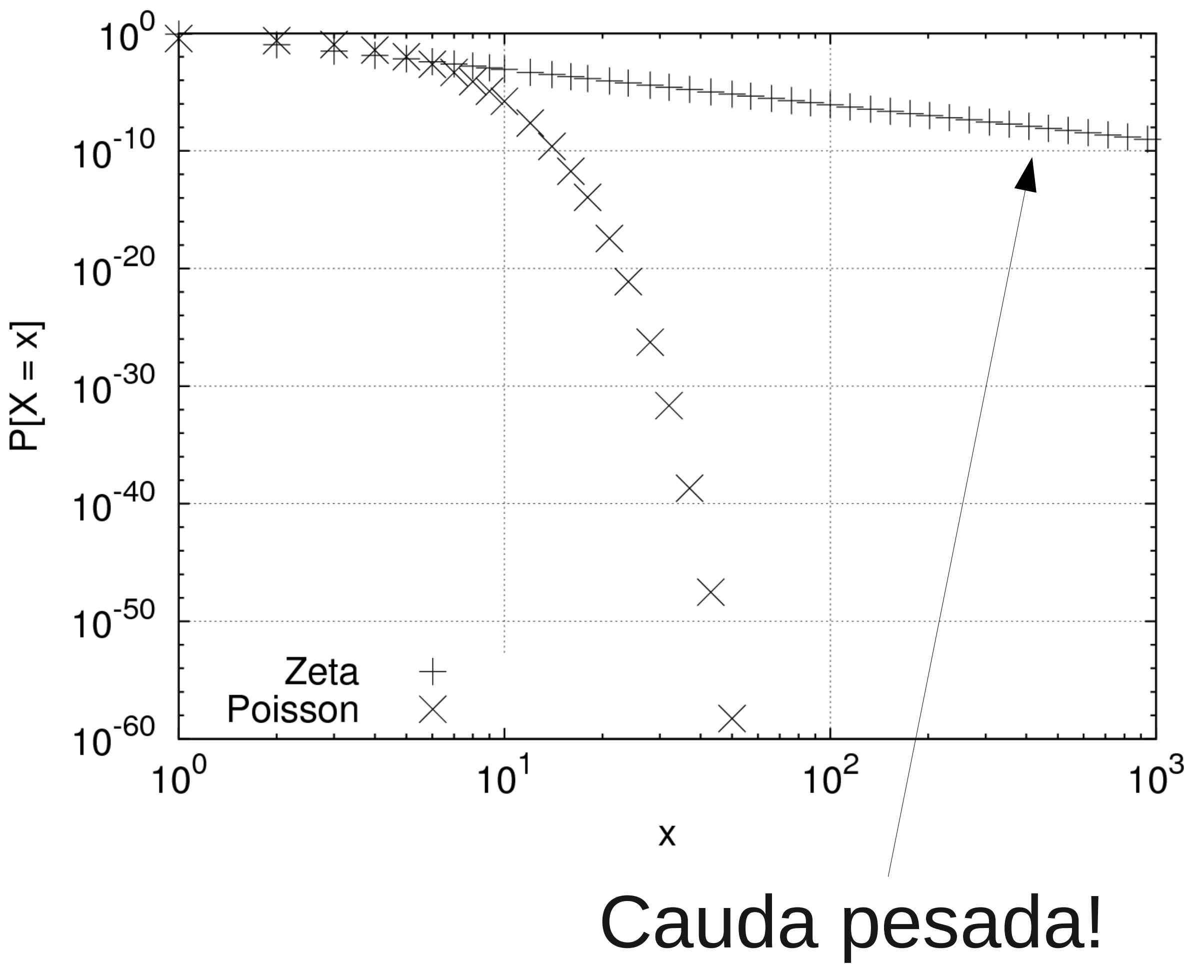
- Temos então

$$\log(f(k)) \sim \log(c\, k^{-a}) = -a \log k + \log c$$

- Relação linear entre log f(k) e log k com inclinação -a
- Expoente determina inclinação da reta
- Invariância na escala
  - inclinação não depende de k

Exemplo

- Comparação entre Zeta (a = 3) e Poisson
- Mesmo valor esperado
- Poisson tem cauda exponencial
- Zeta pode gerar valores 10^{50} mais chance!



Distribuições em Lei de Potência

Zipf

- Variação de distribuição de Zeta
- Valores de X são limitados superiormente
- Momentos sempre existem (para qualquer expoente)

Pareto

- Variável aleatória contínua (Zeta e Zipf são discretas)
- Pode ser ou não limitada superiormente
- Aproximação contínua para fenômenos discretos (ex. grau)

Estudando Redes Reais

- Como estudar de forma generalizada uma rede real?
  - ex. Internet, Facebook

Modelo matemático!

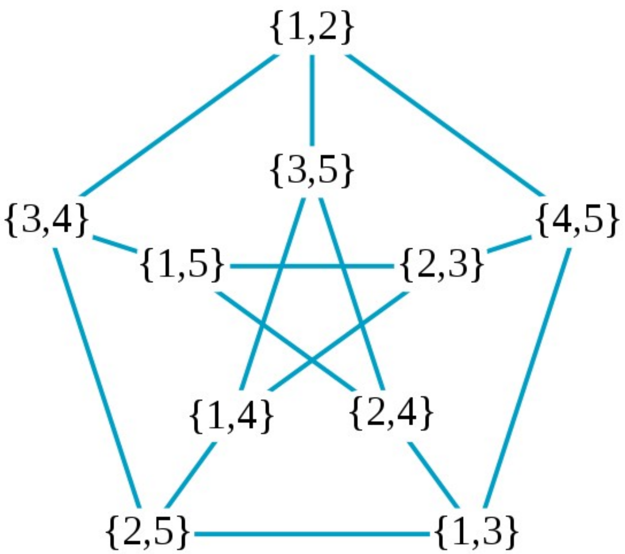
- Abstração matemática da realidade
- Permite trabalhar analiticamente com realidade simplificada
  - vai além do estudo empírico da realidade

Modelo matemático para redes

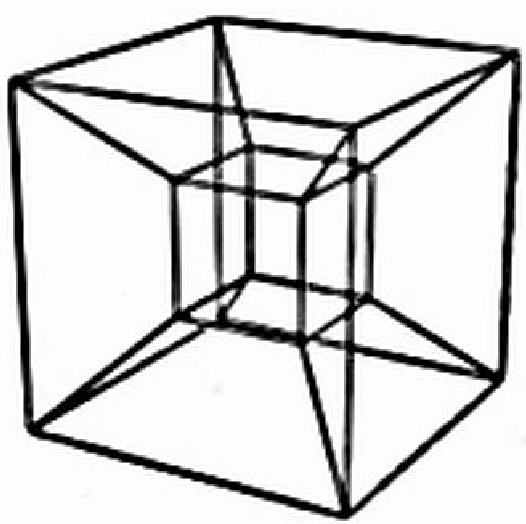
Modelos Determinísticos para Redes

- Estrutura topológica é determinística
- Propriedades e características topológicas são determinísticas

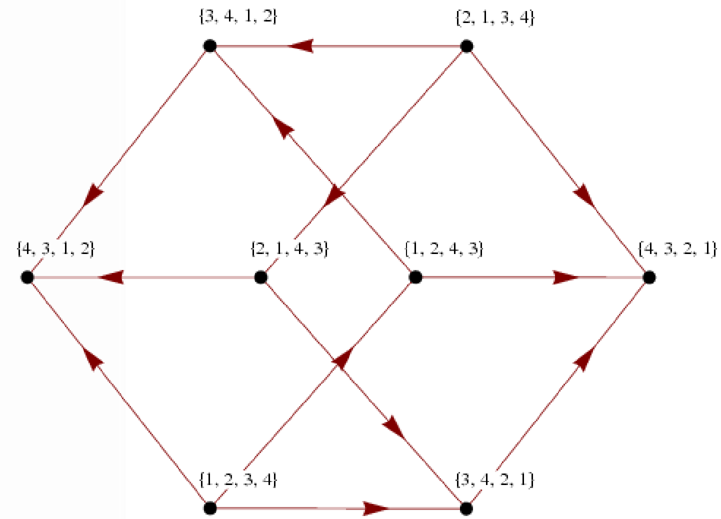
Exemplos de modelos?



Grafo Kneser



Hipercubo

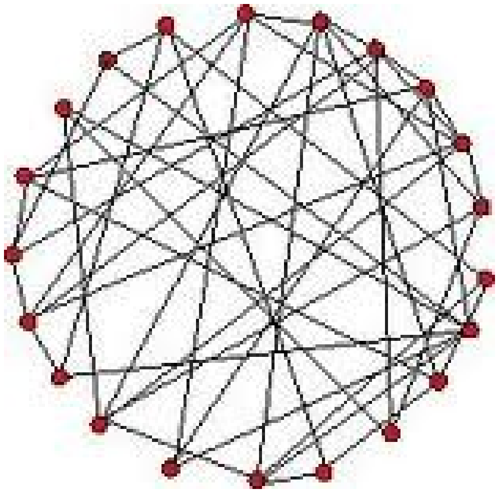


Grafos de Caley

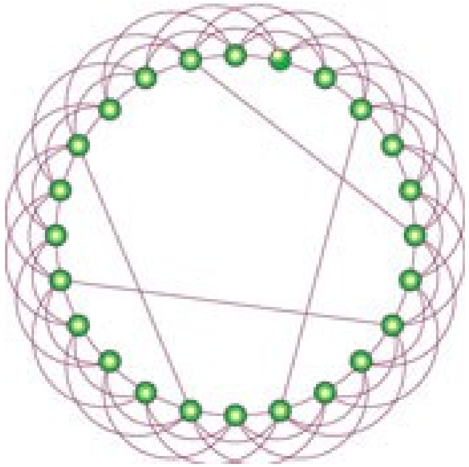
Modelos Probabilísticos para Redes

- Estrutura topológica é aleatória
  - definido através de um processo aleatório

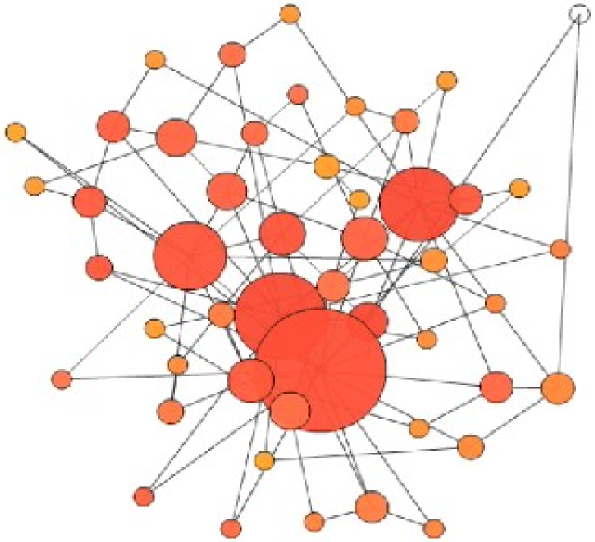
Exemplos de modelos?



Erdös-Réyni



Watts-Strogatz



Barabási-Albert

Modelo G(n,p)

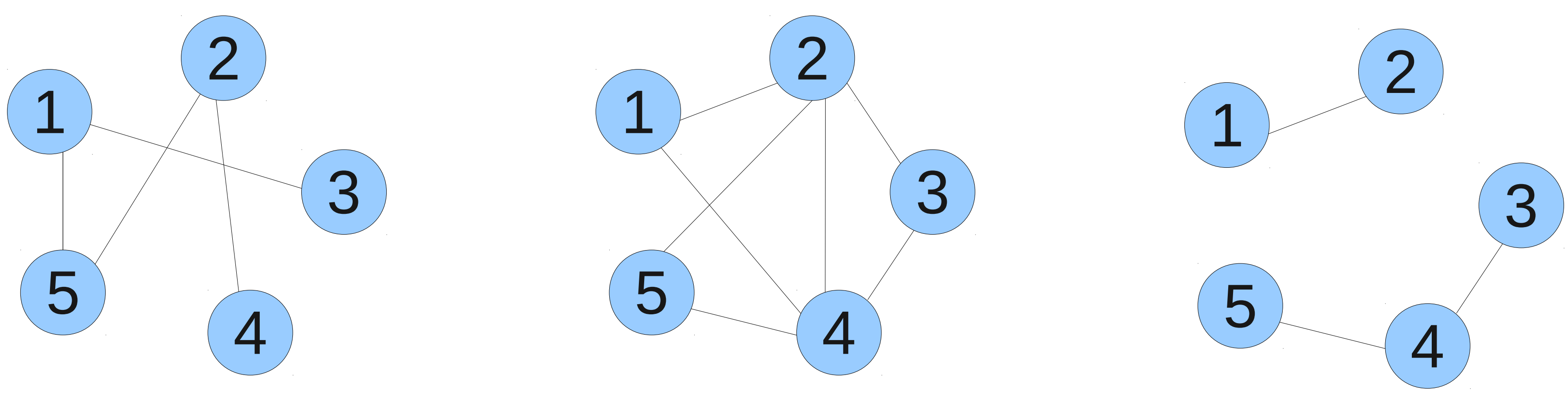
- Modelo clássico para grafos aletórios
  - introduzido por Erdös e Rényi na década de 50
  - aka. modelo binomial, modelo de Erdös-Rényi

Como funciona o modelo?

- Grafo possui *n* vértices rotulados
- Cada possível aresta do grafo existe com probabilidade *p* independentemente
  - grafo não-direcionado, sem loops

Modelo G(n,p)

- Modelo possui dois parâmetros (determinísticos)
  - n: número de vértices
  - p: prob. de existência de cada aresta
- Dado os dois parâmetros, grafo gerado é único?
  - Ex. n=5, p=0.5
- Não! Grafo é aleatório, será amostrado



Estudo do G(n,p)

- Estudar a **estrutura** do modelo G(n,p)
  - em função de n e p (únicos parâmetros)
- Estrutura é aleatória
  - depende da realização
- Estrutura depende de n e p
  - efeitos de borda quando n é pequeno
- Caracterizar estrutura para *n* muito grande
  - estrutura quando *n* tende para infinito

Como caracterizar a estrutura?

Propriedades Simples

- Espaço amostral do modelo G(n,p)?
  - Quantos grafos diferentes?

$$|S| = 2^{\binom{n}{2}}$$

Cada aresta pode ou
não estar presente

- Probabilidade do modelo gerar um grafo G específico com n vértices?
  - Gerar um conjunto de arestas E = {e_1, e_2, ... , e_m }

$$P(gerar\, conjunto\, E) = p^{|E|}(1-p)^{\binom{n}{2}-|E|}$$

- Depende apenas de |E|
- Todos grafos são equiprováveis quando p = 1/2

Arestas do G(n,p)

- Quantas arestas tem um grafo do modelo G(n,p)?
- Variável aleatória, M. Distribuição de M?

$$P(M=m)=\binom{\binom{n}{2}}{m} p^{m}(1-p)^{\binom{n}{2}-m}$$

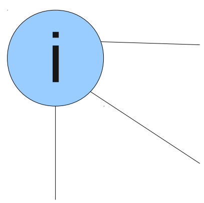
Distribuição binomial

- Valor esperado de M?

$$E[M]=\binom{n}{2} p=\frac{n(n-1)p}{2}$$

Grau do G(n,p)

- Qual é o grau de um vértice qualquer do modelo?
- Variável aleatória D. Distribuição de D?



$$P[D=k]?$$ k = 0, 1, ..., n-1

- Cada aresta incide sobre vértice i com prob p

$$P[D=k]=\binom{n-1}{k}p^k(1-p)^{n-1-k}$$

Distribuição binomial

- Valor esperado do grau?

$$E[D]=(n-1)p$$

Coeficiente de Clusterização

- De um vértice qualquer (escolhido ao acaso)
- Variável aleatória, K. Distribuição? Valor esperado?
- Distribuição condicional, no grau do vértice

$$P\left[K=\frac{k}{\binom{d}{2}}\middle|D=d\right]$$

Prob. de termos k arestas entre os d vizinhos do vértice

- Valor esperado

$$E[K]=p$$

Não depende do grau

Propriedades Topológicas

- Seja propriedade X
  - ex. X = “grafo é conexo”
- Seja G um grafo decorrente da realização do processo aleatório
  - G é uma amostra do modelo

G possui propriedade X?

- Subconjunto do espaço amostral possui propriedade X
- G possui X probabilisticamente

Propriedades Topológicas

- Seja p_G a probabilidade do modelo gerar o grafo G
  - grafo G é determinístico, uma amostra
- Seja C o conjunto de grafos (determinísticos) que possuem propriedade X
- Então temos

$$P[G\ possuir\ X] = \sum_{G' \in C} p_{G'}$$

- Probabilidade pode depender de *n*
  - número de vértices do grafo
- Interesse na probabilidade assintótica
  - quando *n* é muito grande

Propriedade de Quase Todos os Grafos

- Seja G um modelo aleatório para grafos e X uma propriedade topológica
- Se

$$P[G\ possuir\ X] \rightarrow 1 \quad \text{quando} \quad n \rightarrow \infty$$

- Então dizemos que quase todos os grafos de G possuem X
  - X ocorre em G a.a.s. (*asymptotically almost surely*)
- Análogo quando probabilidade tende a zero
  - X nunca ocorre em G a.a.s.

Comportamento de p

- Caracterizar G(n,p) quando n tende a infinito. Mas p?

Como p varia quando $n \rightarrow \infty$

- Duas alternativas
  - p é fixo, não varia
  - p é uma função de n, varia com n
- Influência fundamental na estrutura
  - estrutura depende de p

G(n,p) com p fixo

- Considere $p > 0$ fixo (constante)
- Grau esperado dos vértices?
  - quando n tende a infinito?
- Número esperado de arestas no grafo?
  - quando n tende a infinito?

Grafo muito denso

- **Intuitivamente**: Grafo conexo, vértices muito próximos

G(n,p) com p fixo

- Para p constante > 0, G(n,p) é conexo a.a.s.
- Para p constante > 0, G(n,p) possui diâmetro 2 a.a.s.
- Podemos provar esta propriedades
  - bom exercício
- G(n,p) com p fixo não é muito interessante

G(n,p) com p(n)

- p(n) é uma função de n
  - caso anterior p(n) = cte

Qual p(n)?

- p(n) deve **decrescer** com n
  - Figura!
- Exemplos

$$p(n) = \frac{c}{n} \qquad p(n) = n^{-2.5} \qquad p(n) = \frac{c \log n}{n}$$

G(n,p) com grau médio fixo

- Grau médio fixo, igual a z

$$p(n) = z/(n-1)$$

- Propriedades estruturais dependem de z
  - estudo em função de z
- Tamanho das componentes conexas

Intuição?

- Divisor de águas: $z < 1$, $z > 1$?

Componentes Conexas

- z : grau médio de um vértice
- $z < 1$ (subcrítico)
  - componentes conexas tem tamanho $O(\log n)$ (a.a.s.)
  - Muitas componentes conexas
- $z > 1$ (supercrítico)
  - maior componente conexa tem tamanho $\Omega(n)$
  - todas outras de tamanho $O(\log n)$

Transição de fase na estrutura do grafo!

Adequação do Modelo G(n,p)

- O quão adequado é o modelo G(n,p) para representar redes reais?

Captura número de vértices e grau médio!

- Suficiente?
- Depende dos objetivos...

Captura outros aspectos estruturais das redes reais?

Aspectos Estruturais

- Presente em muitas redes reais
  - Distâncias pequenas
  - Alta clusterização
  - Distribuição do grau segue lei de potência
- Modelo G(n,p)
  - Distâncias: O(log n / log z)
  - Clusterização: p
  - Distribuição do grau: Binomial(n-1, p)

Muito diferentes!

Aspectos fundamentais para muitas aplicações

Exemplo

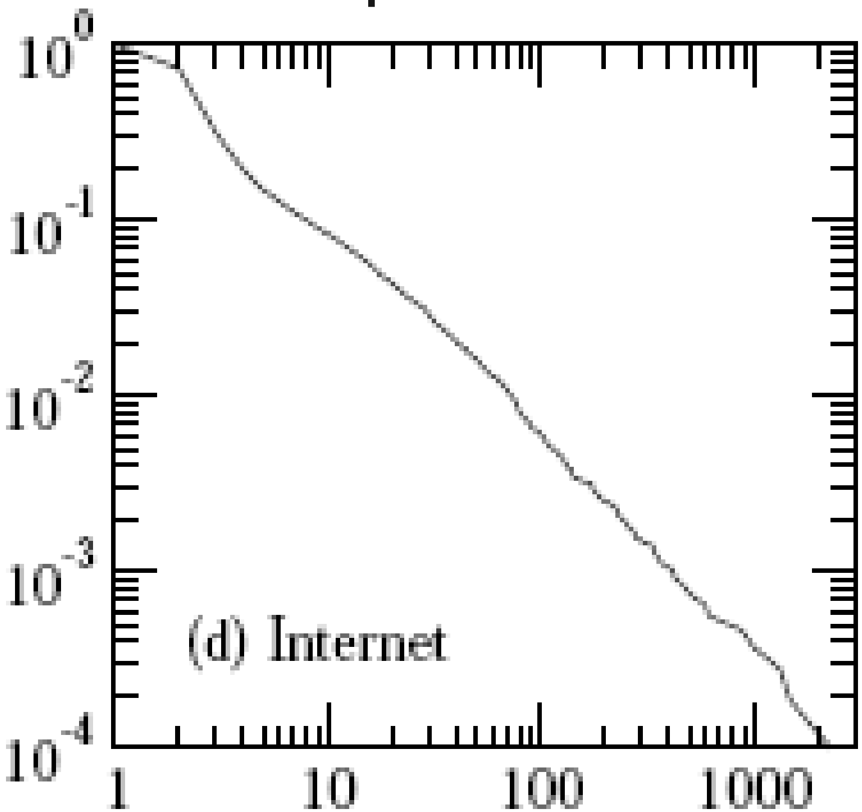
- AS Graph, 10697 vértices e 31992 arestas
- Distância média
  - empírica: 3.31
  - modelo: O(log n/log z) = 5.18
- Clusterização
  - empírica: 0.39
  - modelo: p = 0.00056

~1000 vezes menor!

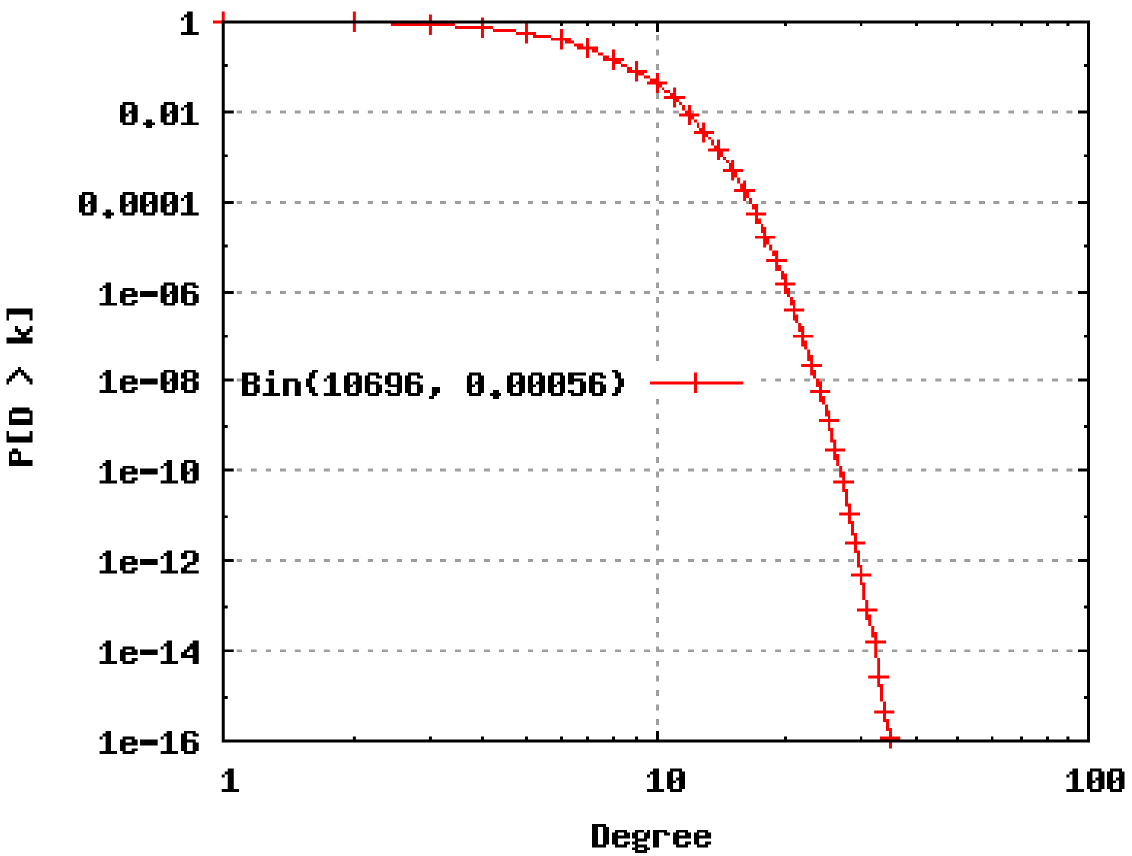
Exemplo

- AS Graph, 10697 vértices e 31992 arestas
- Distribuição do grau: P[D > k]

Empírica



Modelo G(n,p)



Fudamentalmente diferente!

Modelos Generativos

- Explicam o surgimento das propriedades estruturais
  - estrutura é função do processo gerador
- Vértices e arestas são adicionados incrementalmente
  - refletem o “crescimento” da rede
- Processo de crescimento tenta capturar a realidade
  - processo gerador leva a estrutura observada

Muitos modelos propostos são desta classe

Preferential Attachment (PA)

- Regra fundamental (e antiga) de formação
- **Idéia:** objetos novos têm preferência em se relacionar com objetos mais populares

$$P[u \sim v] \approx pop_v$$

prob. de novo objeto u
se relacionar com objeto v

- Objetos populares “atraem” novos objetos
- Idéia possui muitas aplicações
  - objetos e popularidade depende da aplicação

Exemplos ?

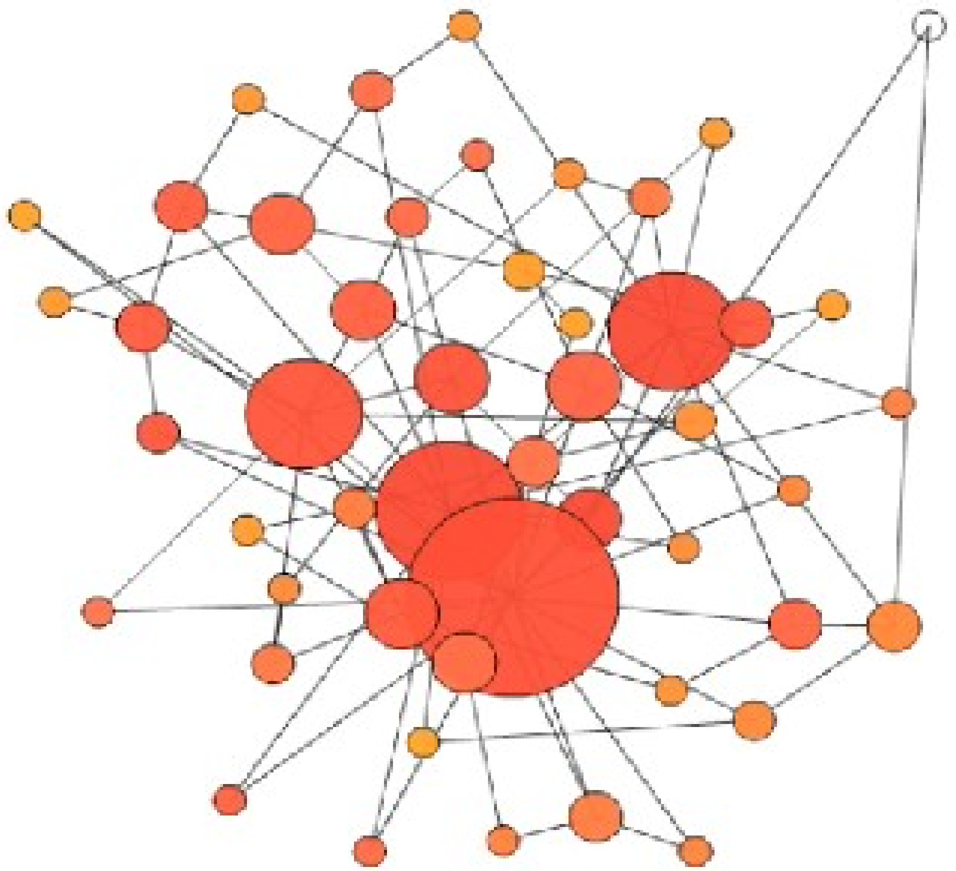
Modelo para Redes

- Modelo para grafos (redes) aleatórias baseado em PA (Barabasi et al 99)
  - objetos são vértices e popularidade é dada pelo grau do vértice
- Processo de formação
  - Inicialmente temos pequeno clique
  - a cada passo, adicionar 1 vértice com grau m
  - vértice adjacente é escolhido aleatoriamente, com prob. proporcional ao seu grau

PA Exemplo

- Clique inicial com 3 vértices, m = 2
- Tamanho do vértice proporcional ao grau

PA Exemplo



- O que está acontecendo?
- "*Rich gets richer*" -> Lei de potência?

Definindo Preferências

- Dado grafo G=(V, E) no instante t = 1, 2, ...
- $d_u(t)$: grau do vértice u no instante t
- Considere a chegada de um novo vértice no instante t
  - vértice traz m arestas
- $p_u(t)$: prob. do vértice u ser incidente a uma nova aresta no instante t

$$p_u(t) = \frac{d_u(t)}{\sum_{v \in V} d_v(t)} = \frac{d_u(t)}{2mt}$$

número de arestas do grafo em t

Evolução do Grau do Vértice

- Quanto vale $d_u(t)$?
  - $d_u(t)$ é uma v. a. discreta
- Assumir $d_u(t)$ é determinístico (valor esperado) e contínuo
- Seja t_u o instante de entrada do vértice u
- Podemos mostrar

$$d_u(t) = m\left(\frac{t}{t_u}\right)^{1/2}$$

- Grau de um vértice cresce como lei de potência
- Mais antigos tem grau maiores

Distribuição do Grau

- Distribuição do grau no tempo t
  - distribuição depende to tempo?

$$P[d_u(t) < k]$$

- Assumir que instante de chegada é uniforme
  - Vértices u entra entre [1, t]
- Podemos mostrar:

$$P[d_u(t) = k] \approx \frac{2m^2}{k^3}$$

Lei de potência na distribuição do grau

Observaçõe sobre PA

- Fenômenos observados podem ser modelados por PA
- Modelo pode levar ao surgimento de *lei de potência* na popularidade dos objetos
- PA como *explicação* para leis de potência observadas empiricamente
  - lei de potência surge, pois processo de formação obedece ao PA
- **Cuidado**: nem toda lei de potência é explicada por PA

Limitação do Modelo BA

- Lei de potência com expoente 3
  - muitas redes tem outros expoentes
- Vértices mais antigos tem grau maiores
  - vértices novos não tem chance de serem populares
- Vértices criam arestas somente no instante de chegada
  - não há evolução local da rede
- Rede gerada tem baixa clusterização

Modelo Small World

- Inspirado nos resultados de Milgram
- Inspirado na alta clusterização observada empiricamente
  - proposto por Watts and Strogatz (Matemática Aplicada) em 1998
  - 30 anos depois de Milgram!
  - Início de “Redes Complexas”
- Modelo captura as duas propriedades
- Mostra que ambas podem coexistir em um mesmo grafo esparso

Clusterização e Distâncias

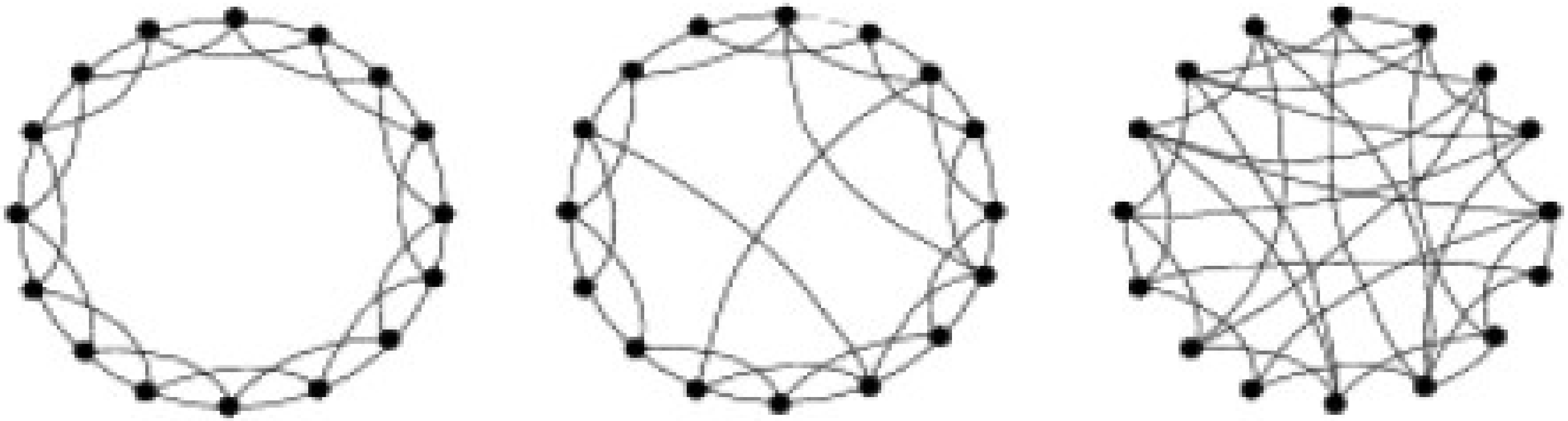
- Considere um grafo esparso
  - número de arestas relativamente pequeno
- Considere alta clusterização e caminhos curtos

Propriedades Antagônicas?

- **Alta clusterização**: muitos triângulos, arestas “desperdiçadas”, caminhos não avançam
- **Caminhos curtos**: arestas usadas eficientemente, avanço mais rápido

Modelo Small World

- Começar com um látice regular
  - N vértices organizados em um círculo
  - arestas para os K vizinhos mais próximos



- Para cada aresta, reposicionar suas extremidades com probabilidade *p*
  - nova extremidade escolhida de maneira uniforme entre os vértices

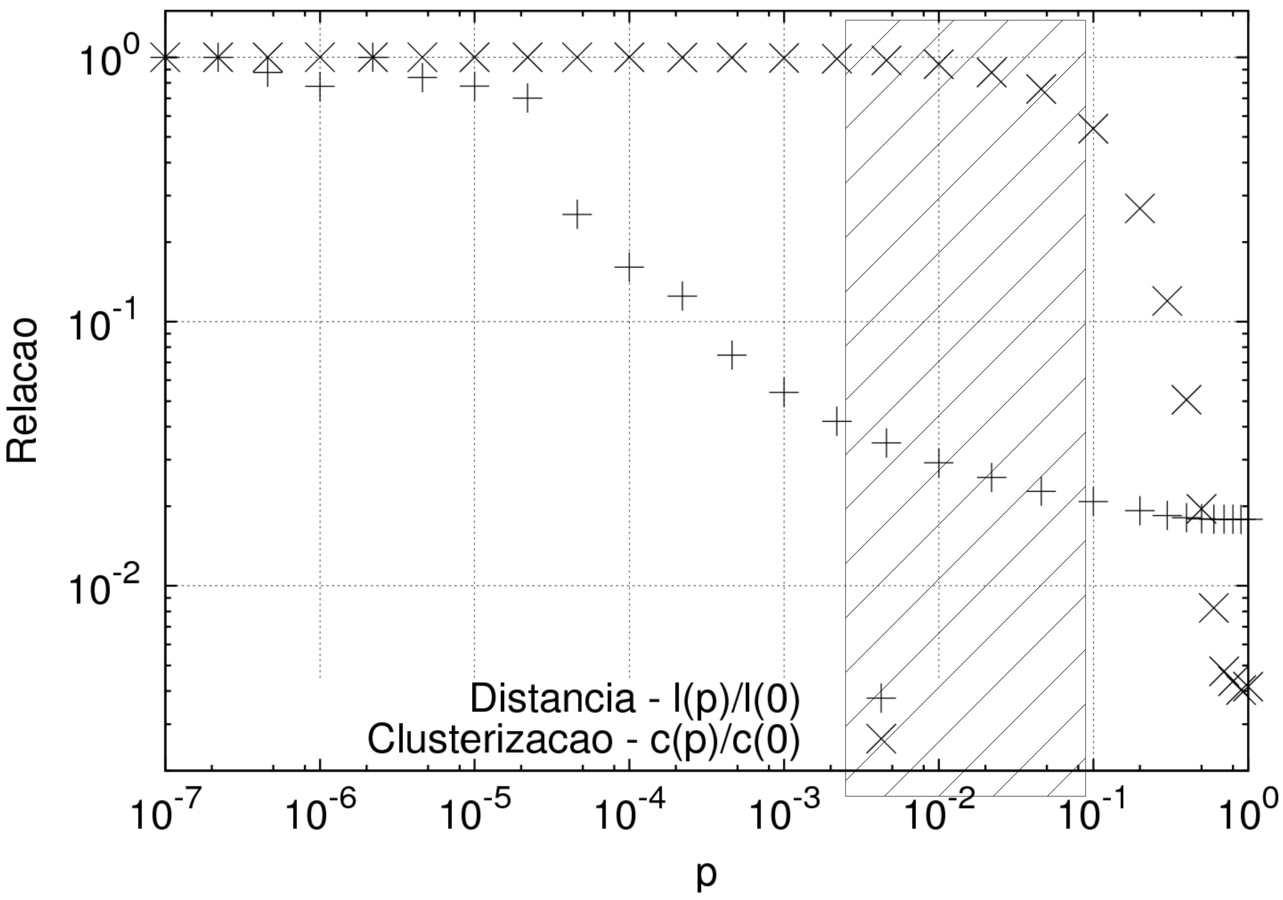
Propriedades Estruturais

- Considere n grande, k grande, com n >> k
- C(p) : clusterização em função de p
  - C(0), alta clusterização (intuitivamente)
  - C(1), baixa clusterização
- l(p) : distância média em função de p
  - l(0), distância média alta (intuitivamente)
  - l(1), distância média baixa

O que ocorre quando *p* é intermediário?

Análise Numérica

- Relação entre C(p)/C(0) e l(p)/l(0)
  - C(0) maior clusterização, l(0) maior distância média



- Distância média decresce rapidamente
- Clusterização decresce mais devagar
- Alta clusterização, ditâncias curtas é possível!
- Small World Networks

Limitação do Modelo SW

- Distribuição de grau não possui cauda pesada
- Não representa um fenômeno real para criação da rede
  - Não possui crescimento

- Muitos outros modelos surgiram na última década
- Modelos que capturam função da rede (e do domínio)
- Problema ainda em aberto