

VERİ MADENCİLİĞİ

Sosyal Ağlar

Prof. Dr. Şule Gündüz Öğüdücü

1

Sosyal Ağ

- Sosyal ağ kişiler arasındaki ilişkilerin oluşturduğu bir yapıdır
- Sosyal ağ incelemesi: ağ yapısının, kişiler ya da gruplar (topluluklar) arasındaki ilişkilerin ve bilgi akışının incelenmesi

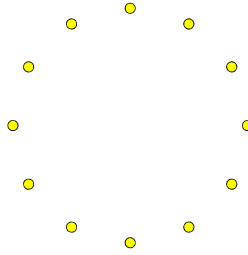


L.C. Freeman, Visualizing Social Networks. Journal of Social Structure, 2000.

2

Sosyal İlişkiler: Çizge

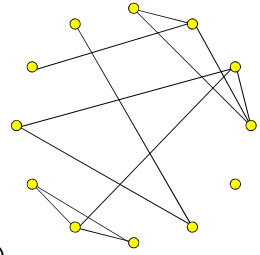
- Düğümmler: Kişiler



3

Sosyal İlişkiler: Çizge

- Düğümmler: Kişiler
- Ayrıtlar: sosyal ilişkiler
 - aile, arkadaş, iş
- Çizge $G(V, E)$
 - V: düğüm kümesi
 - E: Ayrıtlar kümesi
- Benzerlik Matrisi



S. Milgram (1967)

Yakınlığın Altı Derecesi: Six Degrees of Separation

4

Kevin Bacon Oyunu

- 1994 yılında bir grup öğrenci tarafından icat edildi: Craig Fass, Brian Turtle, Mike Ginelly
- Amaç: Bütün oyuncuları en az sayıda bağlantı ile en kısa sürede Kevin Bacon'a bağlamak
- Oracle of Bacon Web sitesi IMDB veritabanındaki veriyi kullanarak iki oyuncu arasındaki en kısa yolu buluyor.
<http://oracleofbacon.org/>



5

Sosyal Ağların Özellikleri

- Farklı (doğal) ağlar
 - sosyal, biyolojik, teknik, içerik..
- Ortak özellikleri
 - Çok büyük, dinamik: düğümler, ayrıtlar eklenebilir/silinebilir
 - düğümmler hangi düğümlerle ilişkide olacaklarına kendileri karar veriyorlar
 - düğümmler arası etkileşim ayrıtlarla sınırlı
 - uzaklık/benzerlik için soyut bilgi: coğrafi, içerik, ilişkiler
- Sosyal ağ kuramı: link analizi
 - Farklı ağların genel özellikleri nelerdir?
 - Bu özellikler nasıl belirlenir, nasıl ölçülür?

6

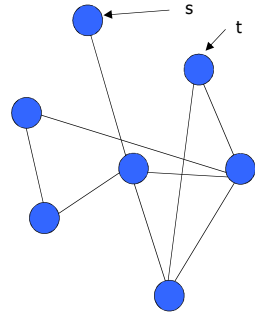
Sosyal Ağların Özellikleri

- Uzunluk ($d_G(s,t)$): İki düğüm (s,t) arasındaki en büyük/ en küçük/ ortalama uzaklık
 - İki düğüm arasında bulunan yoldaki ayrıt sayısı
 - İki düğüm arasında bulunan yoldaki ayrıtların ağırlıklarının toplamı
- Derece: yönlü ise düğüme gelen (in-link) / düğümden çıkan (out-link) bağlantıların sayısı
- Merkez: Ağdaki diğer düğümlerin bağlı olduğu bir ya da bir kaç düğüm
- Yoğunluk: Ağdaki bağlantı sayısının olası bütün bağlantı sayısına oranı

7

Sosyal Ağların Özellikleri

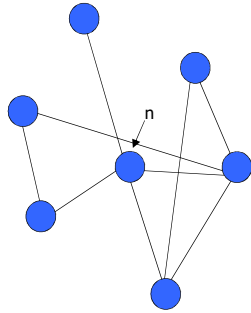
- Uzunluk ($d_G(s,t)$): İki düğüm (s,t) arasındaki en büyük/ en küçük/ ortalama uzaklık
 - İki düğüm arasında bulunan yoldaki ayrıt sayısı
 - $d_G(s,t) = 3$ (en kısa yol)
 - İki düğüm arasında bulunan yoldaki ayrıtların ağırlıklarının toplamı



8

Sosyal Ağların Özellikleri

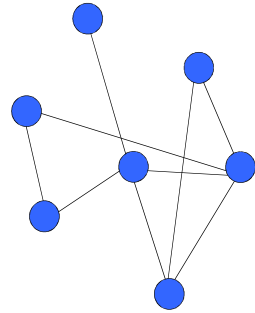
- Derece:
 - yönlü ise düğüme gelen (in-link) / düğümden çıkan (out-link) bağlantıların sayısı
 - yönsüz ise düğüme bağlanmış ayrıt sayısı
 - $d(n)=4$



9

Sosyal Ağların Özellikleri

- Merkez: Ağdaki diğer düğümlerin bağlı olduğu bir ya da bir kaç düğüm
- Yoğunluk: Ağdaki bağlantı sayısının olası bütün bağlantı sayısına oranı
- $\rho=9/21$



10

Tanımlar

- $\sigma_{st} = \sigma_{ts}$ s ve t ($s, t \in V$) düğümleri arasındaki en kısa yol sayısı $\rightarrow \sigma_{ss}=1$
- $\sigma_{st}(v)$: $v \in V$ düğümünün üzerinde bulunduğu s ve t düğümleri arasındaki en kısa yol sayısı

$$C_c(v) = \frac{1}{\sum_{t \in V} d_G(v,t)} \quad \text{closeness centrality (Sabadussi, 1966)}$$

$$C_G(v) = \frac{1}{\max_{t \in V} d_G(v,t)} \quad \text{graph centrality (Hage and Harary, 1995)}$$

$$C_S(v) = \sum_{s \neq v, s \in V} \sigma_{st}(v) \quad \text{stress centrality (Shimbel, 1953)}$$

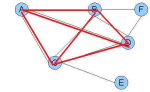
$$C_B(v) = \sum_{s \neq v, s \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad \text{betweenness centrality (Freeman, 1977; Anthonisse, 1971)}$$

Ulrik Brandes, A Faster Algorithm for Betweenness Centrality, Journal of Mathematical Sociology 25(2):163-177, (2001).

11

Tanımlar

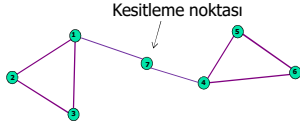
- Hizip (Clique):
 - seçilebilecek her düğüm çifti arasında bir bağ olan alt çizge
 - tam bağlı alt çizge
- Daha zayıflatılmış
 - N-Hizip (N-Clique): Bir düğümün içinde bulunduğu alt çizgedeki diğer tüm düğümlere olan uzaklığı en çok N olabilir
 - N-Klan (N-Clan): N-Hizipteki düğüm çiftleri arasındaki yollar üstündeki düğümler de N-Hizip üyesi
 - K-Plexes: Bir düğümün n düğümden oluşan bir N-Hizip içindeki en az n-k düğüm ile doğrudan bağlı olması



12

Tanımlar

- Kesitleme noktası (Cut Points): Bağlı olan bir G çizgesinden, v düğümü ve bu düğüme bağlı olan bütün ayrıtlar çıkarıldığında oluşan G-v çizgesi bağlı değil ise v kesitleme noktasıdır.



13

Tanımlar

- Kesitleme noktası (Cut Points): Bağlı olan bir G çizgesinden, v düğümü ve bu düğüme bağlı olan bütün ayrıtlar çıkarıldığında oluşan G-v çizgesi bağlı değil ise v kesitleme noktasıdır.



14

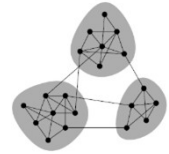
Sosyal Ağların Yeteneği

Özellik	Tanım	Etki
Derece (Degree)	Bir düğümün bağlantı sayısı	daha fazla seçenek
Yakınlık	Diğer düğümlere olan yolun uzunluğu	Diğer düğümlerle doğrudan etkileşim
Ara düğüm	İki düğüm arasında yer alan düğüm	Diğer iki düğüm arasında ilişkiye sağlamak/kesmek

15

Problemler

- Bağlı parçalar:
 - kaç parça, büyüklükleri ne, ne kadar bağlılar
- Ağ çapı:
 - en uzak – ortalama
 - Bağlı olmayan düğümler / parçalar
 - Küçük dünya özelliği
- Demetleme
 - Gruplaşan ilişkiler/düğümmler
 - Örtüşen gruplar
 - Grup içi ve gruplar arası ilişkilerin oranı
 - Grup içi ve gruplar arası ilişkilerin rolü
- Ağın yapısı
 - düğümmlerin derecesi
 - k_i : i düğümünün derecesi, N: i düğümünün komşular kümesi
 - ilişkilerin ağırlığının dağılımı
 - ağ içinde önemli rolü olan düğümler: iki grubu birbirine bağlayan

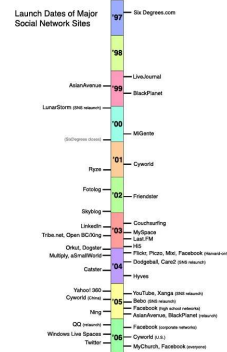


16

Web'de Sosyal Ağlar

- Facebook
 - Myspace
 - LinkedIn
 - Classmates
 - Orkut
 - Bebo
- Medya paylaşım siteleri:
- YouTube
 - Flickr

17



Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), article 11. http://jcmc.indiana.edu/vol13/issue1/boyd_ellison.html

18

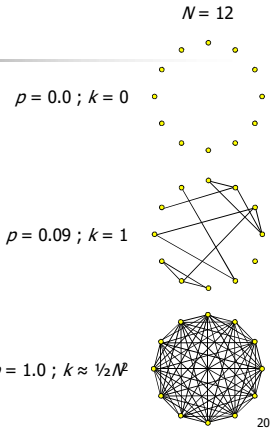
Sosyal Ağlar için Modeller

- Rassal çizgeler (Random Graphs: Erdős-Rényi models)
- Watts-Strogatz modelleri
- Scale-free Networks

19

Rassal Ağlar

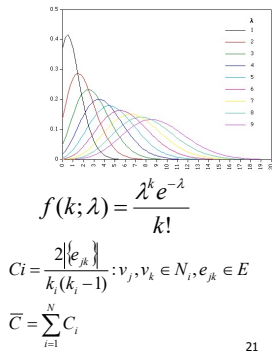
- Erdős-Rényi (ER) Model: 1959 yılında Paul Erdős ve Alfred Rényi
- N: Düğüm sayısı
- p: iki düğüm arasında ayrıt olma olasılığı
 - $p = 1/2N$, $p = 1/N$, $p = 2/N$, $p = 10/N$, $p = \log(N)/N...$
- Ortalama derece: $k \approx pN$



20

Rassal Ağlar

- $N(N-1)/2$ adet hileli yazı tura atma
- Derece dağılımı
 - N çok büyük olduğunda Poisson dağılımı
 - $G(N, p)$ ile bir çizge oluştur, rastgele bir u düğümü
 - $\Pr[\deg(u) = k]$?
 - Poisson dağılımı ortalama $\lambda = p(N-1) \sim pN$



- Demetleme katsayısı (clustering coefficient) küçük

21

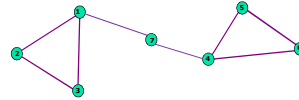
Demetleme Katsayısı

- Demetleme katsayısı

$$C_7 = 2 \cdot 0 / 2 \cdot 1 = 0$$

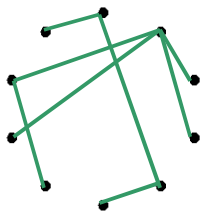
$$C_1 = 2 \cdot 1 / 3 \cdot 2 = 1/3$$

$$C_2 = 2 \cdot 1 / 2 \cdot 1 = 1$$



22

Erdős-Rényi Model (1960)

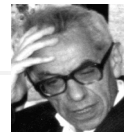


Connect with probability p

$$p = 1/6$$

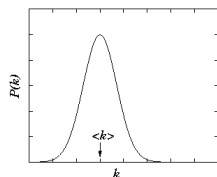
$$N = 10$$

$$\langle \lambda \rangle \sim 1.5$$



Pál Erdős (1913-1996)

Poisson distribution



Erdős Sayısı

- Paul Erdős ile birlikte makale yazma uzaklığı
 - Paul Erdős ile birlikte makale yazan kişinin Erdős sayısı=1
 - Paul Erdős ile birlikte makale yazan bir kişiyle birlikte makale yazan kişinin Erdős sayısı=2
- Yaklaşık 1500'den fazla makalesi var



<http://www.oakland.edu/enp/>

24

Watts-Strogatz modelleri: Caveman and Solaria

- Erdos-Renyi
 - Ortak komşuları olması iki düğüm arasında ayırt olma olasılığını artırmıyor
 - her ayırt daha öncekilerden bağımsız olarak oluşuyor
- Gerçekte oluşan ağ yapısına uygun değil
 - iki kişinin tanışma olasılığı ortak arkadaşları varsa daha fazladır
 - Web de iki sayfa birbirine bağlı ise büyük olasılıkla aynı konudadırlar
- Watts Caveman:
 - ayırtların genel olarak yoğunluğu az
 - ayırtların ortak komşuları varsa aralarında ayırt olma olasılığı büyük
- Watts Solaria
 - ayırtların genel olarak yoğunluğu az, bir düğümün komşuları arasında ayırt olma olasılığı farklı değil
 - Erdos-Renyi çizgesine benzer

25

α -model

- Gerçek hayatta çizgelerde düğümler arasındaki bağlantılar düzenli çizge (regular graph) ve rassal çizge arasında
- Gerçek hayattaki sosyal ağların yapısını modellemek
 - α -model (Watts, D. J. (1999) Kevin Bacon, the small-world, and why it all matters. *Santa Fe Institute Bulletin*, 14(2): center section)
- α -model için parametreler
 - N düğüm sayısı
 - k: ortalama derece
 - p: iki düğüm arasında ayırt olma olasılığı
 - α : yakın ilişkilerin olasılığını artırmak için parametre
 - Rassal ağlarda olduğu gibi iki düğüm arasında ayırt eklenir. Ancak iki düğümün ortak komşuları varsa aralarında ayırt olma olasılığı fazladır.

26

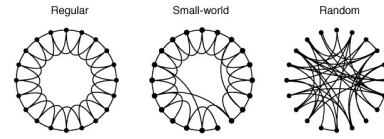
α -model

- u,v düğüm çifti için
 - $m(u,v)$: ortak komşu sayısı
- $R(u,v)$: iki düğümün arasında ayırt olma eğilimi (propensity)
 - $m(u,v) \geq k$, $R(u,v) = 1$
 - $m(u,v) = 0$, $R(u,v) = p$
 - diğer durumlarda $R(u,v) = p + (m(u,v)/k)^\alpha (1-p)$
- $\alpha \rightarrow \infty$ için Erdos-Renyi çizgelerine benziyor

27

Watts-Strogatz Model

- Bir daire etrafında eşit dağılmış N düğüm
- Her düğümün en yakın k komşusu arasında k ayırt (yakın ilişki)
- p olasılığı ile bir düğüme az sayıda rastgele ayırt ekle (uzak ilişki)
- farklı p değerleri için farklı çizgeler



$p = 0$ ———— Increasing randomness ———— $p = 1$

Collective dynamics of 'small-world' networks
Duncan J. Watts & Steven H. Strogatz

28

Small Worlds - Occam's Razor

- küçük α değerleri için demetleme katsayısı büyük
 - Erdos-Renyi $\rightarrow$ çapı küçük
 - α -model $\rightarrow$ büyük demetleme katsayısı
- Occam's Razor
 - farklı özellikler için basit tek bir model
- Watt's small world:
 - çapı küçük
 - demetleme katsayısı büyük

29

Örnek

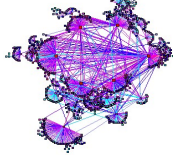
- üç gerçek ağ üzerinde inceleme
 - Oyuncular
 - batı bölgesindeki güç santralleri
 - C.elegans sinir sistemi

	n	k	d	c
Actors	225,226	61	3.65	0.79
Power-grid	4,941	2.67	18.7	0.08
C.elegans	282	14	2.65	0.28

30

Small World

- Rastgele iki düğüm arasındaki yolun uzunluğu kısa
- Serbest Ölçekli Ağlar (Scale free):
 - Derece dağılımı: Power Law $P(k) \sim k^{-\gamma}$
- Örnek web
 - düğüm web sayfaları
 - ayrıntlar bağlantılar



31

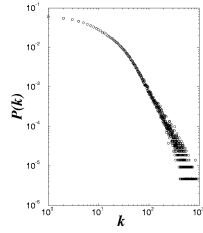
Serbest Ölçekli Ağlar

- Düğüm sayısı N sabit değil
 - Ağ sürekli yeni düğümlerin eklenmesi/silinmesi ile değişiyor
 - web: yeni sayfalar/siteler ekleniyor/siliniyor
 - yayınlar: yeni yayınlar ekleniyor
- Eklenen düğümlerle oluşan ayrıntlar üniform değil
 - çok sayıda sayfanın/sitenin bağlantı verdiği sayfaya/siteye bağlantı verme olasılığı yüksek
 - çok sayıda yayının referans gösterdiği yayının referans gösterilme olasılığı yüksek

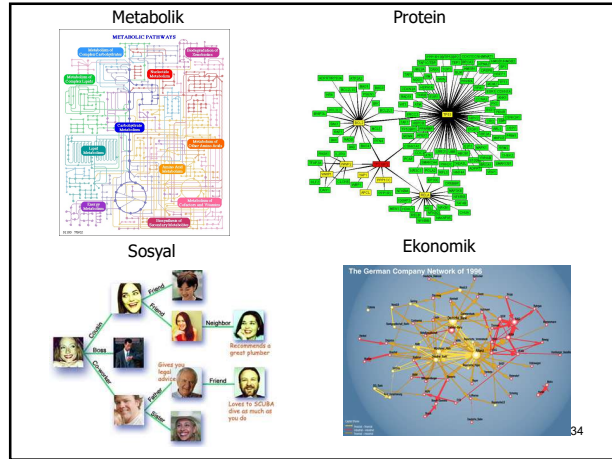
32

Serbest Ölçekli Ağlar

- Pareto veya power law dağılımı
 - $P(k) \sim k^{-\gamma}$
 - çapı küçük ($\sim \log(N)$)
- demetleme katsayısı çok büyük değil
 - yakın komşuları ile ayrıt olma olasılığı yüksek değil
 - Merkez ("hub") olan düğümlerle ayrıt oluşturma olasılığı yüksek



33



34

Çizge içindeki Gruplar

- Bölmek istediğimiz grup sayısının belirli olup olmamasına bağlı olarak iki problem:
 - Çizge bölmeleme (graph partitioning)
 - Topluluk belirleme
- Çözülmesi gereken problem genel olarak aynı
 - Her problem için farklı algoritmalar

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

35

Çizge Bölmeleme

- Çizge bölmelemede çizgenin kaç alt gruba (k) bölüneceği belirlenir
- Çizge büyüklüğü belirlenen k alt çizgeye bölünür
 - Farklı altçizgelerde yer alan düğümler arası ayrıntların sayısını (ağırlığını) en küçültme
 - Alt çizgelerin büyüklüğü yaklaşık olarak belirlenebilir
 - çizgenin alt çizgelerde yer alan düğüm sayısı eşit olacak şekilde iki alt çizgeye bölünmesi

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

36

Topluluk belirleme

- Topluluk belirlemede grup sayısı veya büyüklüğü belirlenmez
- Amaç ağı doğal olarak ayrıldığı hatları bulmaktır.
 - Grup içinde ayrıt sayısının fazla olması, gruplar arasında ayrıt sayısının az olması
- Topluluk belirleme çizge bölmeleme kadar net tanımlanmış bir problem değildir
 - Grup içi ve gruplar arası ayrıtların oranı için farklı ölçütler

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

37

Çizge Bölmeleme/Topluluk Belirleme

- Çizge bölmeleme büyük bir çizgeyi işlenebilir alt parçalara bölmek için kullanılabilir
- Topluluk belirleme büyük bir çizgenin yapısını anlamak için kullanılabilir
 - Bağlanma yapısını anlamak için
- Çizge parçalama sonunda belirlenen sayıda ve özellikte alt çizgeler elde edilebilir
- Topluluk belirleme sonunda doğal topluluklar yoksa topluluk elde edilmeyebilir

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

38

Topluluk Belirleme

- S. Fortunato and C. Castellano, *Community Structure in Graphs*, ArXiv e-prints
- Bir çizge içinde ortak özellikleri/görevleri olan düğümler topluluğu
 - aynı konudaki web sayfaları
 - benzer işleve sahip proteinler
 - aynı konuda çalışan insan grupları
 - aynı ilgi alanına sahip insan grupları
- Topluluk içindeki topolojik konumlarına göre düğümler sınıflandırılabilir
 - topluluğun merkezinde yer alan düğüm
 - topluluğun sınırında yer alan düğüm

39

Topluluk Tanımı

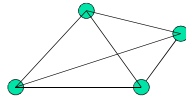
- Topluluklar çizge yapısına göre tanımlanır
- En yaygın tanım: topluluk içindeki ayrıtların yoğunluğu daha fazladır
 - Dar tanım: Sadece alt çizgedeki ilişkilere göre toplulukları belirliyor.
 - Düğüm tabanlı
 - Grup tabanlı
 - Geniş tanım: Tüm çizgenin yapısı inceleniyor
 - Düğüm benzerliği

40

Dar Tanım

- Düğüm Tabanlı

Örnek: Hızip, n-klan, k-plexes



- Hesaplama maliyeti yüksek
- Sosyal ağlarda uygulanması kısıtlı

41

Dar Tanım

- Grup tabanlı
- Her grup belirli kriterleri sağlamalı
 - örnek: grup yoğunluğu bir eşik değerinden büyük olmalı
- $G_s = (V_s, E_s)$ alt çizgesi γ - dense quasi clique
$$\frac{2|E_s|}{|V_s|(|V_s| - 1)} \geq \gamma$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

42

Geniş Tanım

- Tüm çizgenin yapısı göz önünde bulundurularak topluluk belirleme
- Topluluk çizgenin yapısal bir birimi
 - Farklı alt çizgeler arasındaki ayrıtların sayısının/ağırlığının az olması
- Altçizgeler hem kendi içindeki ilişkiler hem de çizgenin geri kalanıyla olan ilişkileri ile belirleniyor
- Null model: içinde topluluk bulunmayan çizge
 - Erdős-Rényi
 - Newman-Girvan: düğümlerin orjinal çizge ile aynı dereceye sahip olduğu rassal çizge

43

Düğüm benzerliği

- Düğümler arası benzerlik hesaplama
 - Farklı benzerlik hesaplama yöntemleri
- Düğümler birbirine "benzer" ise aynı toplulukta
 - dar
 - geniş

44

Düğüm Benzerliği

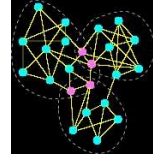
- Veri kümesi $D=\{x_1, x_2, \dots, x_n\} \Rightarrow G(V,E)$
 - ε -komşu çizge: Veri kümesindeki nesneler arasında uzaklık hesaplanır. Aralarındaki uzaklık belirlenen ε değerinden küçük olanlar arasında ayrıt oluşturulur.
 - k -en yakın komşu çizge: Veri kümesindeki nesnelerin k en yakın komşuları arasında ayrıtlar oluşturulur.
 - Tam bağlı çizge: Bütün nesneler arasında benzerlik hesaplanır ve bu benzerlik ile ağırlıklandırılmış ayrıtlar oluşturulur.

$$s(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

45

Örtüşen Topluluklar

- Gerçek dünyada bir nesne birden fazla gruba üye olabilir

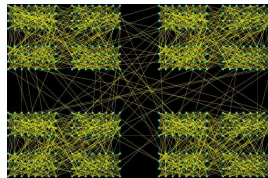


G. Palla, I. Derényi, I. Farkas, T. Vicsek,
*Uncovering the overlapping community structure of
complex networks in nature and society*
Nature 435, 814, 2005
<http://www.cfinder.org/>

46

Hiyerarşik Yapı

- Altçizgeler tekrar parçalanabilir

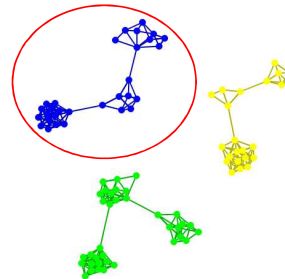


A. Clauset, C. Moore, M.E.J. Newman,
Hierarchical structure and prediction of missing links
Nature 453, 98, 2008

47

Topluluk Belirleme

- Hangisi daha iyi



48

İki Farklı Sonuç

- Üstünlüklerini karşılaştırmak için bir kriter Q
- $Q(P1) > Q(P2)$ veya $Q(P1) < Q(P2)$?

49

Birimsellik

- Modularity:

$$\text{Bitişiklik Matrisi: } A_{uv} = \begin{cases} 1 & (u,v) \in E \\ 0 & \text{diğer} \end{cases}$$

$$m = \sum_{u,v \in V} \frac{A_{uv}}{2}$$

$$k_u = \sum_{v \in V} A_{uv}$$

$$C = \{c_1, c_2, \dots\}, c_i \cap c_j = \emptyset (i \neq j) \cup_{c_i \in C} c_i = V$$

$$e_{ij} = \sum_{u \in c_i, v \in c_j} A_{uv} / 2m$$

$$a_i = \sum_{v \in c_i} k_v / 2m$$

$$Q(G, C) = \sum (e_{ii} - a_i^2)$$

- Büyük modularity değeri iyi bir topluluk yapısı

50

Girvan-Newman Algoritması

- M. Girvan & M.E.J Newman, *Community structure in social and biological networks*, PNAS 99, 7821-7826 (2002)
- Toplulukları birbirine bağlayan ayrıtları belirle
 - Betweenness
- bu ayrıtları silme
 - birbirinden ayrık demetler

51

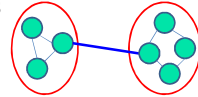
Edge Betweenness

- Bir ayrıtlın ağırlığı edge betweenness değeri ile ölçülebilir:

Ayrıt üzerinden geçen en kısa yol sayısı

$e(1,2)$ ayrıtının edge betweenness değeri $4=6/2+1$
2 ve $\{4, 5, 6, 7, 8, 9\}$ düğümleri arasındaki yol $e(1,2)$ veya $e(2,3)$ üzerinden geçmesi gerekiyor $(6/2)$
 $e(1,2)$ 1 ve 2 düğümü arasındaki yolda $(+1)$

- Daha büyük edge betweenness değeri olan ayrıtlar iki topluluk arasında köprü



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

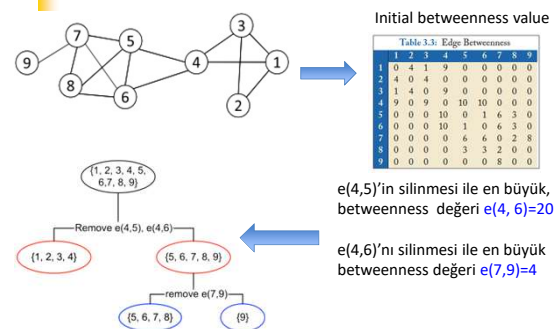
52

Girvan-Newman Algoritması

- Bütün ayrıtların "betweenness" değerleri hesaplanır
- En büyük "betweenness" değerine sahip ayrıtlar silinir
- Kalan ayrıtların "betweenness" değerleri hesaplanır
2. adıma geri dönülür

53

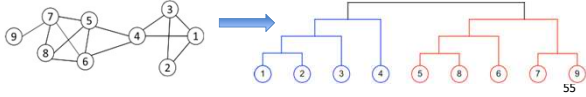
Edge Betweenness değerine göre Bölünmeli Demetleme



54

Geliştirilmiş Girvan-Newman Algoritması

- M.E.J. Newman & M. Girvan, *Finding and evaluating community structure in networks*, Phys. Rev. E 69, 026113 (2004)
- 1. Her düğüm bir demet
- 2. Q değerini en büyütecek iki düğümü birleştir
- 3. Bütün düğümler tek demet olana kadar işleme devam et
- 4. En fazla Q değerine sahip demetlemeyi seç



Problemler

- Örtüşen topluluklar
- Yönlü çizgeler
- Ağırlıklı çizgeler
- Karmaşıklık
- Dinamik ağlar
- Büyük ağlar

56