Boğaz'da Yapay Öğrenme
İsmail Arı Yaz Okulu
2-5 Temmuz 2018

Doku ve Hastalıklara Özgü Büyük Ölçekli Biyolojik Ağların Oluşturulması ve Analizi

Tolga Can

Bilgisayar Mühendisliği Bölümü

ODTÜ

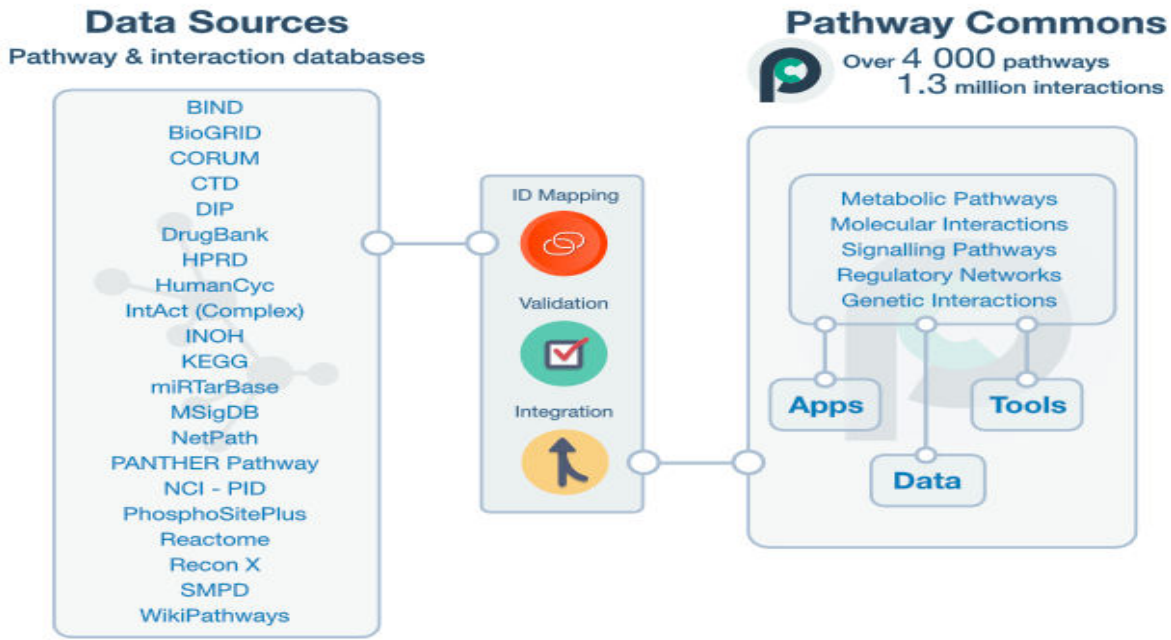
İçerik

- Genom ölçeğinde biyolojik etkileşim verileri
  - Pathway Commons portalı
- Doku ya da hastalıklara özgü biyolojik etkileşim ağlarının oluşturulması
- Çizgecikler (İng. Graphlets)
  - Büyük bir çizge içinde çizgeciklerin sayılması
- Büyük çizgelerde yakınlık sorguları
  - Çizgede yeniden başlamalı rastgele yürüme (İng. Random Walks with Restarts)

Pathway Commons

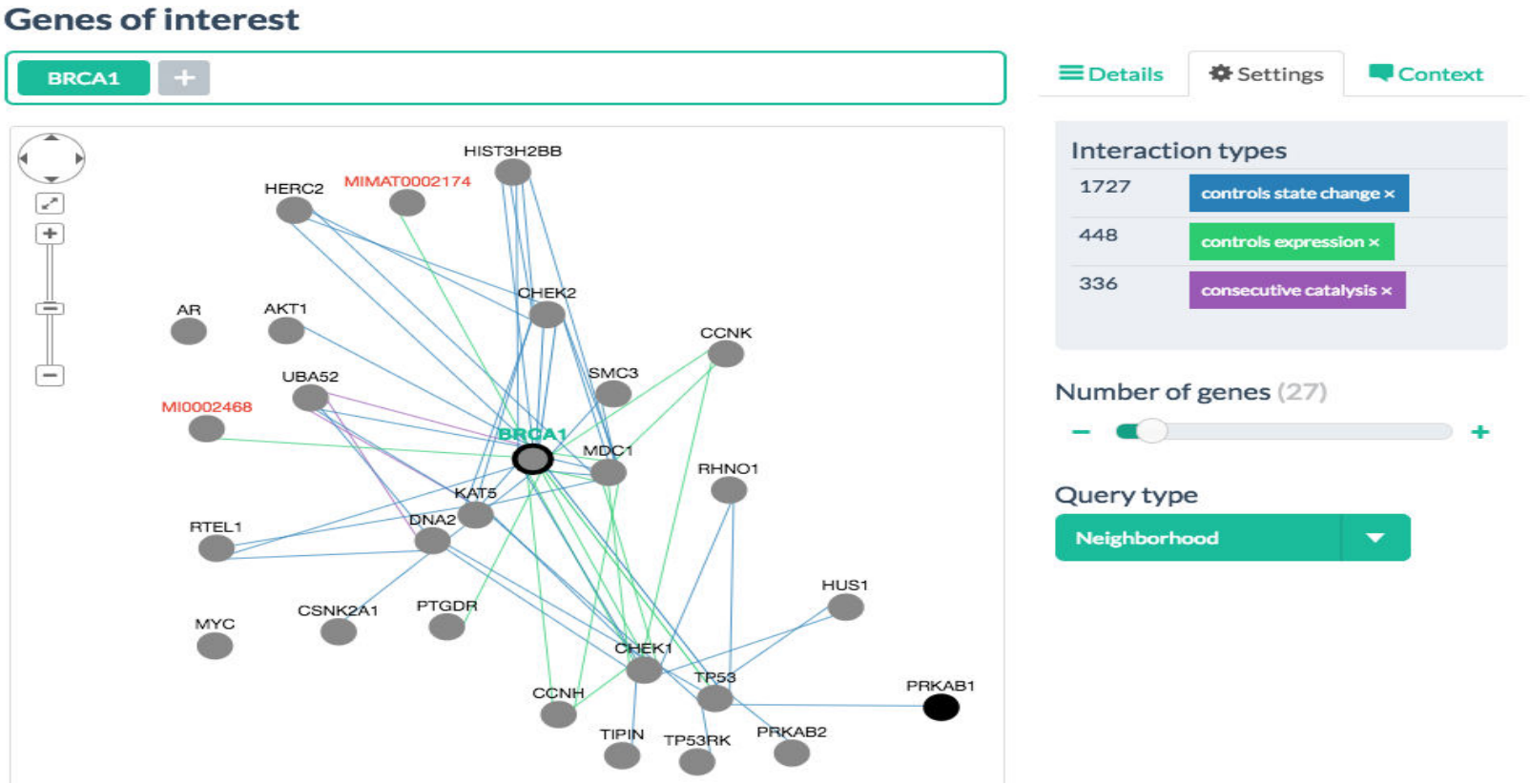
Biyolojik etkileşim verilerini ve bunların analizi için faydalı araçları içeren bir web portalı.

pathwaycommons.org



Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., … Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. Nucleic Acids Research, 39(Database issue), D685–D690.

Pathway Commons Araçlarından Bir Örnek



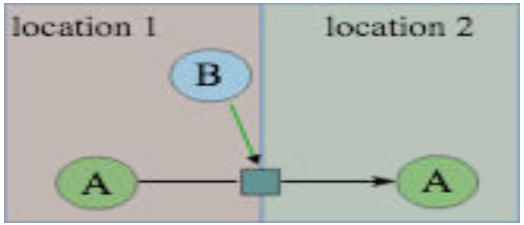
Pathway Commons Versiyon 10 (7 Mayıs 2018)

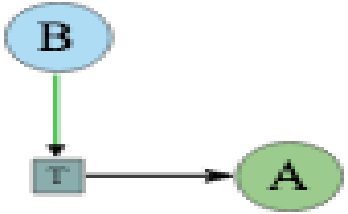
```
A1BG    controls-expression-of  A2M
A1BG    interacts-with  ABCC6
A1BG    interacts-with  ACE2
A1BG    interacts-with  ADAM10
A1BG    interacts-with  ADAM17
A1BG    interacts-with  ADAM9
A1BG    interacts-with  AGO1
A1BG    controls-phosphorylation-of     AKT1
A1BG    controls-state-change-of        AKT1
A1BG    interacts-with  ANXA7
A1BG    interacts-with  CDKN1A
A1BG    interacts-with  CRISP3
A1BG    interacts-with  CRK
A1BG    interacts-with  CSE1L
A1BG    interacts-with  CUL4B
A1BG    interacts-with  DDX3X
A1BG    interacts-with  DEAF1
A1BG    interacts-with  E2F1
A1BG    interacts-with  E2F2
A1BG    interacts-with  E2F3
A1BG    interacts-with  EIF3E
A1BG    interacts-with  ELAVL1
A1BG    interacts-with  FDXR
"PathwayCommons10.All.hgnc.sif" 2374707L, 78865441C
```

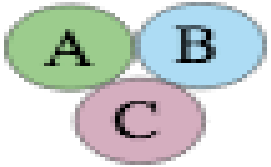
- http://www.pathwaycommons.org/archives/PC2/v10/PathwayCommons10.All.hgnc.sif.gz
- 32,875 protein/küçük molekül arasında 2,374,707 etkileşim

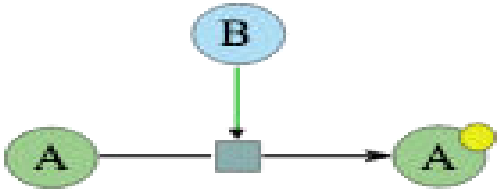
Pathway Commons'taki Etkileşimler

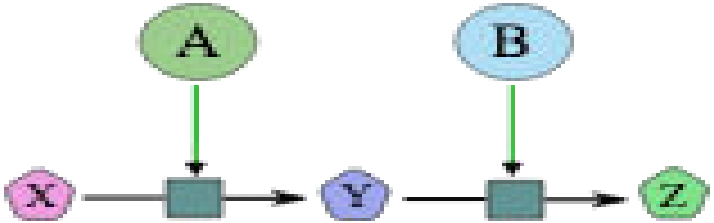
- http://www.pathwaycommons.org/pc2/formats











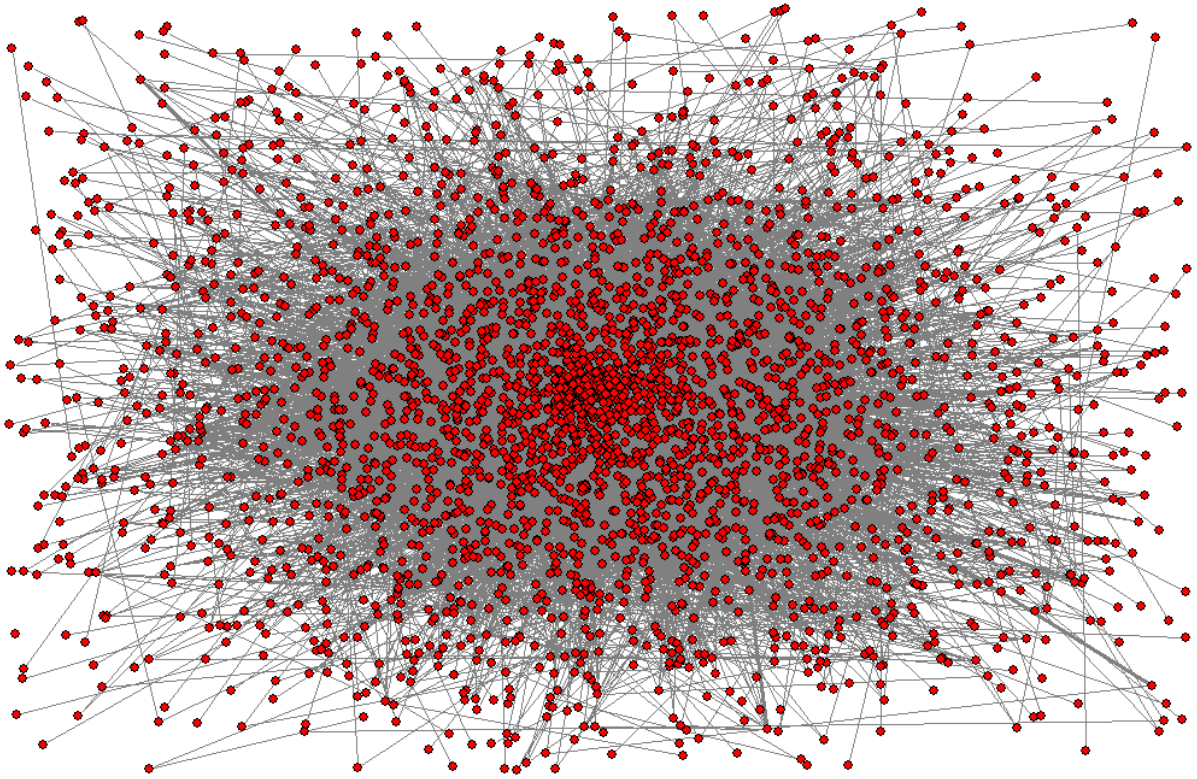
Pathway Commons Versiyon 10 (7 Mayıs 2018)

- 18,931 protein arasında 1,225,798 etkileşim

Etkileşim Türü	Sayısı
interacts-with	434,790
in-complex-with	189,421
controls-state-change-of	250,730
catalysis-precedes	190,578
controls-expression-of	139,194
controls-transport-of	7,776
controls-phosphorylation-of	13,309

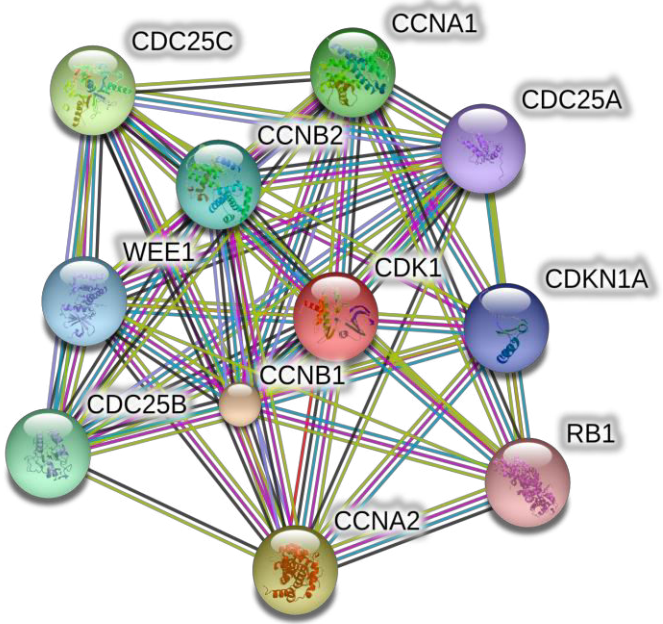
yönlü etkileşimler

Ağın tamamını görüp de anlamak çok zor



ProNet
(Asthana et al. 2004)
S. Cerevisiae
3,112 düğüm
12,594 kenar

Farklı tür etkileşimler içeren bir ağ



string.embl.de

Doku/Hastalıklara Özgü Ağların elde edilmesi

- En doğru yolu: etkileşim belirleme deneylerinin yapıldığı doku/hastalık bilgilerini kullanmak
  - Pathway Commons'daki veri kaynaklarının çoğunda böyle bir bilgi yok ne yazık ki
  - Yüksek ölçekli etkileşim belirleme yöntemleri doku ya da hastalığa özel örneklerde çalıştırılmamış olabiliyor

Doku/Hastalıklara Özgü Ağların elde edilmesi

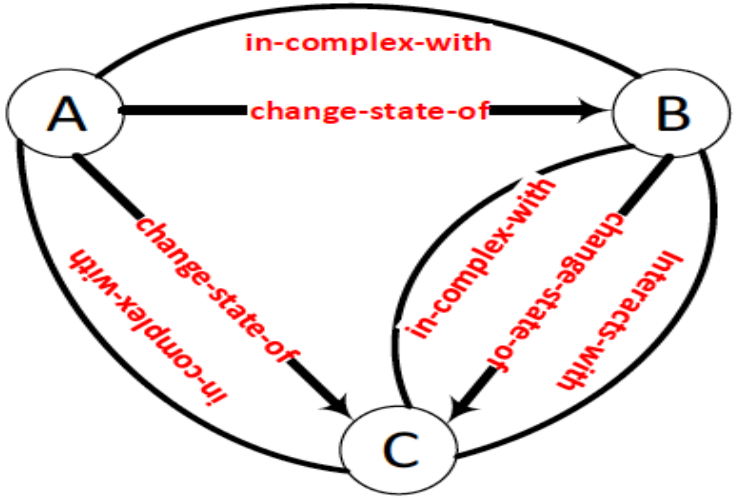
- Daha az doğru ama bir dokuda olmayan etkileşimleri çıkarmanın bir yolu: Transkriptom profilleri
  - NCBI GEO GSE7307: Expression profiling by array
    - Herkese açık veri, April 09, 2007
    - Normal ve hasta insan dokuları Affymetrix U133 plus 2.0 ile profillenmiş
  - Toplam 677 örnek, 90'dan fazla doku tipi
  - 141 farklı hastalık/doku ağı elde etmekte kullanılabilir

Doku/Hastalıklara Özgü Ağlar

- İfade edilmiş genlerin belirlenmesi
  - Basit bir eşik değer kullanımı: ifade değeri 10.0'dan büyük olan bütün genlerin ifade edildiğini varsaymak
- Daha doğru çözümler:
  - Her genin davranışını bütün dokuları göz önüne alıp ayrı ayrı modellemek
  - Referans genler kullanarak normalizasyon yapmak

Çizgeciklerin sayılması

- 2-3 düğümlü çizgeciklerin sayılması için bir yöntem:
  - Her bir 2-3 düğümlü kenar kodlaması Hashtable veriyapısı ile sayılabilir
  - Ama izomorfik olan çizgeciklerin yalnızca bir kez sayılması için dikkatli olmak gerekir



Kenarların kodlanması

A'dan B'ye olan kenar:
1 1 1 0 1 0 1 0

B'den A'ya olan kenar:
1 1 0 1 0 1 0 1

Bu iki kodlamanın izomorfik olduğu gerçeği ile sayım yapmalıyız

İstatistiksel Anlamlılığın Ölçülmesi

- Gerçek biyolojik ağlardaki çizgeciklerin sayılarını rastgele oluşturulmuş (örn. kenar karıştırması yöntemi ile) ağlardaki çizgecik sayıları ile karşılaştırabiliriz. Eğer rastgele ağlarda bu sayılar normal bir dağılım gösteriyorsa:

$$z_g = \frac{c_g - \mu_g}{\sigma_g}$$

Biyolojik Ağlarda Yakınlık Sorguları

- Yakınlık (İng. Proximity) sorguları genlerin fonksiyonlarının belirlenmesi, hastalık-gen ilişkileri, aynı yolakta görev alan genlerin belirlenmesi gibi bir çok amaçla kullanılmaktadır.

Çizgelerde Rastgele Yürüme

- Google PageRank ölçütü de bu yöntemi kullanarak geliştirilmiştir.

Google PageRank

- Varsayım: A sayfasından B sayfasına olan bir bağlantı A sayfasının B sayfasını önerdiğini göstermektedir

➔Yani bir sayfa ne kadar fazla sayfa tarafından öneriliyorsa o kadar önemlidir (PageRank'i fazladır) diyebiliriz

➔Gelen bağlantı sayısının direkt kullanılmasının problemi ne olabilir?

Google PageRank

- Özyineli bir çözüm: Bir sayfanın önemi
  - hem kendisine yapılan bağlantıların sayısıyla,
  - hem de bu bağlantıları yapan sayfaların önemiyle orantılıdır.
  - S. Brin and L. Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." Computer Networks 30:1-7 (1998), 107-117.

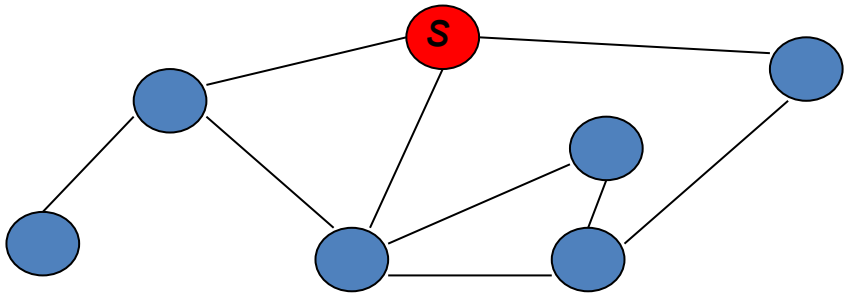
PageRank tanımı

- Sonsuza kadar devam edecek şekilde, aşağıdaki gibi rastgele hareket eden bir web gezgini düşünelim:
  - Başlangıçta gezgin rastgele bir sayfadadır
  - Her adımda gezgin
    - c olasılığı ile tamamen rastgele bir sayfaya
    - ya da 1- c olasılığı ile bulunduğu sayfadaki rastgele bir bağlantıya

    gider
- **PageRank ölçütü limitte gezginin bir *p* sayfasını ne oranda ziyaret ettiğini gösterir.**

Yeniden başlamalı rastgele yürüme

- s düğümünde başlayan bir rastgele yürüme düşünelim. Her adımda bulunduğu düğümün komşularından birine rastgele gidebiliriz ya da s düğümüne c olasılığı ile geri dönebiliriz.



Bir yakınlık ölçütü

- $p_s(v)^{(t)}$ olasılığını rastgele gezginin v düğümünde t. adımda bulunma olasılığı olarak tanımlarsak.
- Limitte değişmez olan $p_s(v)$ değeri v düğümümün s düğümüne olan yakınlığı (İng. affinity) ile orantılıdır ve yinelemeli matris işlemleri ile kolayca hesaplanabilir.

Yakınlık vektörü p'nin durgun halini hesaplama

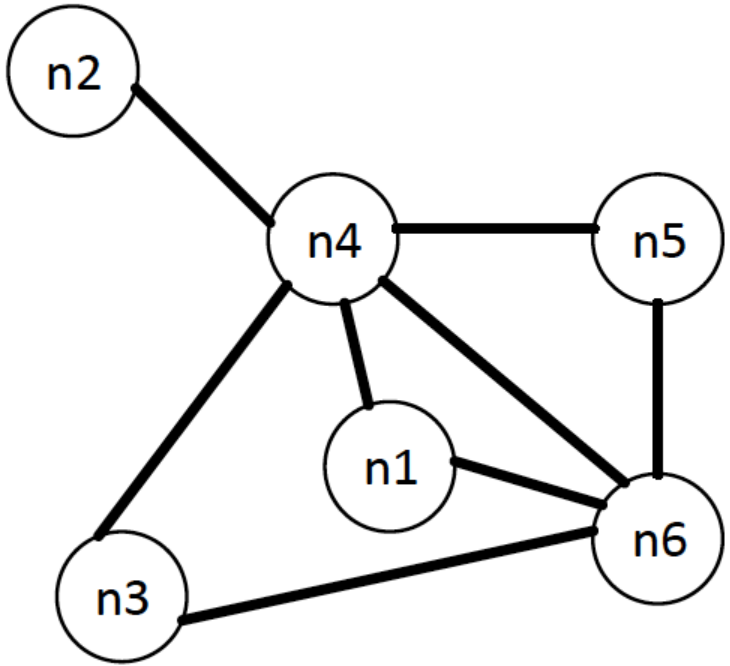
- $\mathbf{s}$ başlangış düğümlerini belirten bir sütun vektörü olsun (yani, $\mathbf{s}_i = 1/n$, eğer i düğümü n tane başlangıç düğümünden biri ise, değilse $\mathbf{s}_i = 0$).
- Aşağıdaki yinelemeli denklemi $\mathbf{p}$ sütun vektörü değişmeyene kadar uygula:

$$\mathbf{p}_{t+1} = (1-c)\mathbf{A}^{\mathbf{T}}\mathbf{p}_t + c\mathbf{s}$$

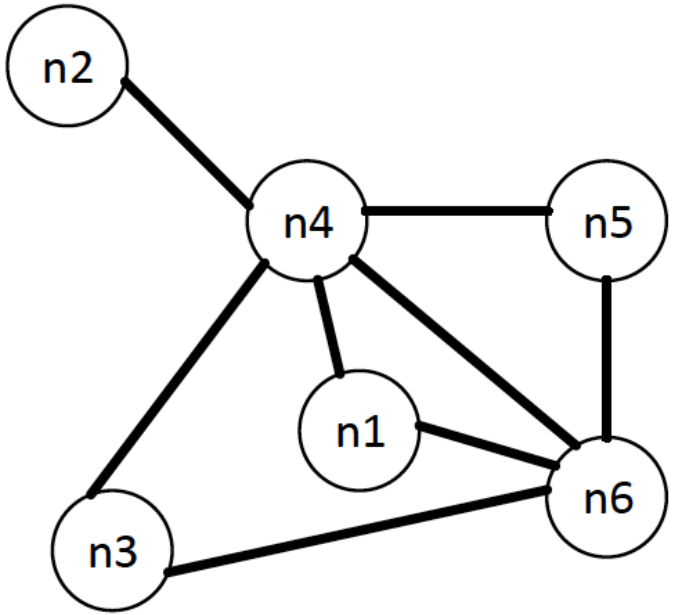
burada $\mathbf{A}$ normalize edilmiş komşuluk matrisini ve c de başa dönme olasılığını göstermektedir.

Küçük bir örnek

- Başlangıç (geri dönüş düğümleri): n_5 ve n_6 olsun.



Yakınlık matrisi ve başlangıç vektörleri



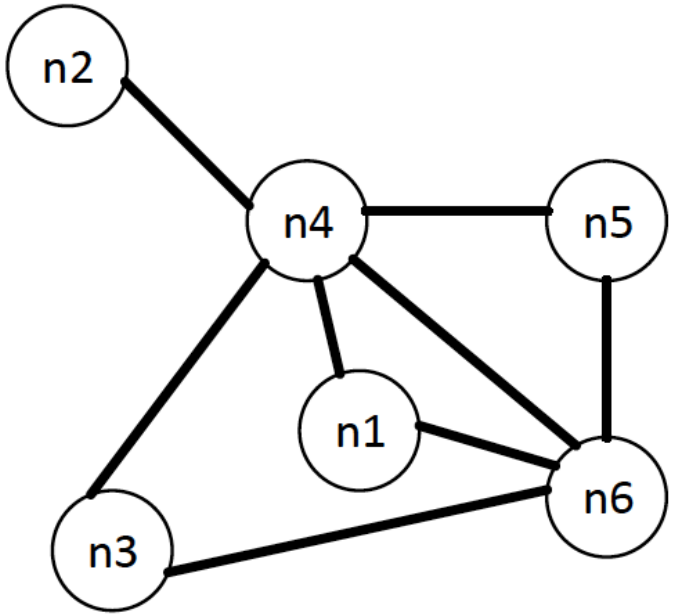
A:

	n1	n2	n3	n4	n5	n6
n1	0	0	0	1	0	1
n2	0	0	0	1	0	0
n3	0	0	0	1	0	1
n4	1	1	1	0	1	1
n5	0	0	0	1	0	1
n6	1	0	1	1	1	0

$s = p_0$

n1	0
n2	0
n3	0
n4	0
n5	0.5
n6	0.5

Normalize edilmiş yakınlık matrisi



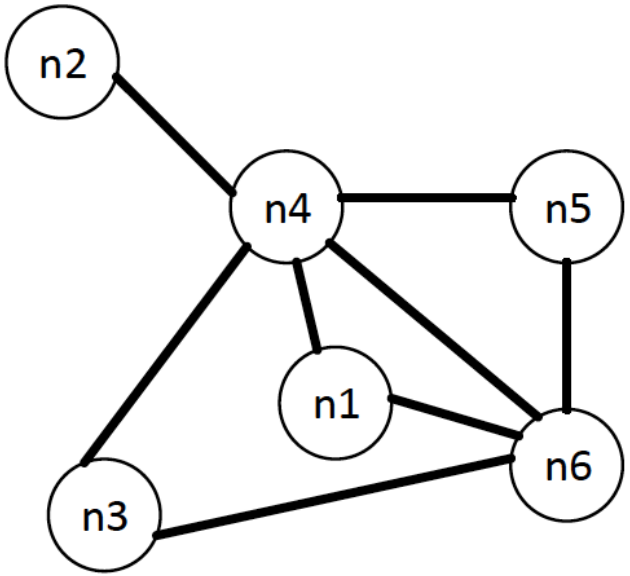
A:

	n1	n2	n3	n4	n5	n6
n1	0	0	0	.5	0	.5
n2	0	0	0	1	0	0
n3	0	0	0	.5	0	.5
n4	.2	.2	.2	0	.2	.2
n5	0	0	0	.5	0	.5
n6	.25	0	.25	.25	.25	0

$s = p_0$

n1	0
n2	0
n3	0
n4	0
n5	0.5
n6	0.5

p_1'in hesaplanması



c = 0.3 olsun

$p_1 = 0.7$

A^T:

	n1	n2	n3	n4	n5	n6
n1	0	0	0	.2	0	.25
n2	0	0	0	.2	0	0
n3	0	0	0	.2	0	.25
n4	.5	1	.5	0	.5	.25
n5	0	0	0	.2	0	.25
n6	.5	0	.5	.2	.5	0

x

p_0:

n1	0
n2	0
n3	0
n4	0
n5	0.5
n6	0.5

+ 0.3

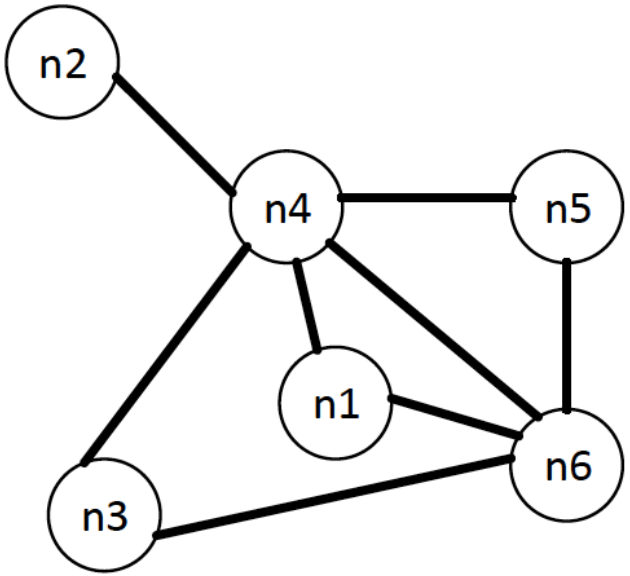
s:

n1	0
n2	0
n3	0
n4	0
n5	0.5
n6	0.5

p_1 =

n1	0.087
n2	0.0
n3	0.087
n4	0.262
n5	0.238
n6	0.325

p_2'nin hesaplanması



$p_2 = 0.7$

A^T:

	n1	n2	n3	n4	n5	n6
n1	0	0	0	.2	0	.25
n2	0	0	0	.2	0	0
n3	0	0	0	.2	0	.25
n4	.5	1	.5	0	.5	.25
n5	0	0	0	.2	0	.25
n6	.5	0	.5	.2	.5	0

x

p_1:

n1	0.087
n2	0.0
n3	0.087
n4	0.262
n5	0.238
n6	0.325

$+ 0.3$

s

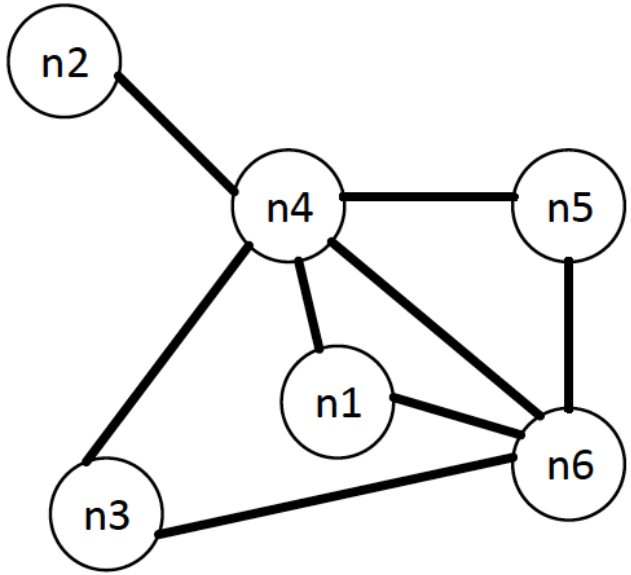
n1	0
n2	0
n3	0
n4	0
n5	0.5
n6	0.5

$p_2 =$

n1	0.094
n2	0.037
n3	0.094
n4	0.201
n5	0.244
n6	0.331

$p_{21} = p_{22}$

n1 0.089
n2 0.032
n3 0.089
n4 0.225
n5 0.239
n6 0.327



Sorular

- Yakınlık sorguları için entegre ağları nasıl sayısal olarak modelleyebiliriz?
- Farklı tür ağlarda yapılmış yakınlık sorgu sonuçlarını nasıl birleştirebiliriz?
- Farklı doku/hastalık ağlarındaki yakınlık sorgu sonuçlarını nasıl karşılaştırabiliriz?

Teşekkürler

- Çizgecik Sayımı: Doktora öğrencim Arzu Burçak Sönmez (ODTÜ Enformatik Enstitüsü Sağlık Bilişimi)
- Ağlarda yakınlık sorguları: Kathy Marcopol ve Ambuj K. Singh (UC Santa Barbara)
- Finansal destekler:
  - TÜBİTAK
    - #144E111 no'lu araştırma projesi
  - Bilim Akademisi Genç Bilim İnsanları Ödül Programı (BAGEP) 2014-2016