

**T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

SOSYAL AĞLARDA HASTALIK İLAÇ BAĞLANTI TAHMİNİ

YÜKSEK LİSANS TEZİ

Müh. İhsan TUĞAL

(111129111)

Anabilim Dalı: Bilgisayar Mühendisliği

Programı: Kuramsal Temeller

Danışman: Doç. Dr. Mehmet KAYA

Tezin Enstitüye Verildiği Tarih: 05 Ağustos 2013

AĞUSTOS – 2013

SOSYAL AĞLARDA HASTALIK İLAÇ BAĞLANTI TAHMİNİ

Müh. İhsan TUĞAL

Yüksek Lisans Tezi

Bilgisayar Mühendisliği Kuramsal Temeller Anabilim Dalı

Danışman: Doç. Dr. Mehmet KAYA

AĞUSTOS - 2013

**T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

SOSYAL AĞLARDA HASTALIK İLAÇ BAĞLANTI TAHMİNİ

YÜKSEK LİSANS TEZİ

Müh. İhsan TUĞAL

(111129111)

Tezin Verildiği Tarih: 05 Ağustos 2013
Tezin Savunulduğu Tarih: 20 Ağustos 2013

Tez Danışmanı : Doç. Dr. Mehmet KAYA (F.Ü.)
Diğer Jüri Üyeleri : Doç. Dr. Mustafa TÜRK (F.Ü.)
Doç. Dr. A. Bedri ÖZER (F.Ü.)

AĞUSTOS-2013

ÖNSÖZ

Sosyal ağların sağlık alanında kullanılmasının getirileri düşünüldüğünde, yapılacak çalışmaların önemi ortaya çıkmaktadır. Bu konudaki çalışmaların geliştirilerek devam ettirilmesi gerekmektedir.

Bu konuda beni yönlendiren ve bu tezin hazırlanması aşamasında bana büyük destek veren danışman hocam Doç. Dr. Mehmet KAYA'ya çok teşekkür ederim. Ayrıca bu tez çalışması için proje desteği sağlayan FÜBAP'a teşekkürlerimi sunarım.

Eğitim hayatım boyunca benden maddi ve manevi desteklerini hiç esirgemeyen aileme sonsuz teşekkürlerimi sunarım.

İhsan TUĞAL

ELAZIĞ - 2013

İÇİNDEKİLER

	<u>Sayfa No</u>
ÖNSÖZ.....	II
İÇİNDEKİLER.....	III
ÖZET.....	IV
SUMMARY.....	V
ŞEKİLLER LİSTESİ.....	VI
TABLolar LİSTESİ.....	VII
KISALTMALAR LİSTESİ.....	VIII
1. GİRİŞ.....	1
2. SOSYAL AĞLAR VE ANALİZİ.....	3
2.1. Çizge Teorisi.....	10
2.2. Sosyal Ağ Analizi Yazılımları.....	12
2.3. Sosyal Ağ Analizindeki Aşamalar.....	15
2.4. Sosyal Ağ Analizindeki Ölçütler.....	15
2.5. Sosyal Ağ Yapılı Verilerin Sınıflandırılması.....	19
3. BAĞLANTI TAHMİNİ.....	20
3.1. Bağlantı Tahmini Problemi.....	21
3.2. Matematiksel Tanımı.....	22
3.3. Yapılan Bazı Çalışmalar.....	23
3.4. İlişki ve Ağ Türleri.....	24
3.5. Benzerlik İlişkileri ve İki-mod Ağlar.....	26
3.6. Bağlantı Tahmini Yöntemleri.....	28
3.6.1. Yerel Benzerlik İndeksleri.....	30
3.6.2. Genel Benzerlik İndeksleri.....	33
3.6.3. Yarı Yerel İndeksler.....	35
4. UYGULAMA SONUÇLARI.....	36
5. SONUÇ.....	48
KAYNAKLAR.....	50
ÖZGEÇMİŞ	

ÖZET

Ağ yapısında olan çeşitli verilerden yeni veriler elde etme ve geleceğe dönük tahminler yapılabilir. Ağ yapısındaki düğümler ve bu düğümlerin aralarındaki ilişkilerden yola çıkılarak, ağın gelecekteki yapısı ve bireylerin oluşturabileceği yeni ilişkilerin veya vazgeçeceği ilişkilerin tahmini yapılmaya çalışılır.

Bu tezde özellikle sosyal ağlarda bağlantı tahmininin tanımı, özellikleri, bu konuda yapılmış çalışmalar irdelendi. Bağlantı tahmini ile ilgili önerilen yöntemlere geçilmeden önce veri madenciliği, sosyal ağlar, sosyal ağ analizi konularında bilgiler verildi.

Uygulama olarak, verilen bağlantı tahmini yöntemi kullanılarak yapay oluşturulan hastalık ilaç ağı üzerinde analizler yapıldı. Bu yöntemlerin kullanılabilirliği, bağlantı tahminin bu ağlara uygulanabilirliği gösterildi.

Anahtar Kelimeler: Sosyal Ağ Analizi, Bağlantı Tahmini, Bağlantı Madenciliği, Veri Madenciliği, Bilgi Keşfi, Hastalık İlaç Ağı

SUMMARY

Through the network structure we can get new data from various data and is intended to generate future predictions. Based on the nodes and relationships between nodes, the future structure of the network and new relationships or give up relationships can be estimated.

The definition, features and studies of link prediction in social networks are examined in this thesis. Before starting to this topic data mining, social networks, social network analysis are explained.

Our application makes analysis on drug disease network by some link prediction methods. The availability of these methods and applicability of link prediction to these networks is shown.

Key Words: Social Network Analysis, Link Prediction, Link Mining, Data Mining, Knowledge Discovery, Drug Disease Network

ŞEKİLLER LİSTESİ

Sayfa No:

Şekil 2.1. Yeni nesil web	6
Şekil 2.2. Sosyal ağ sınıfları	7
Şekil 2.3. Düğümler arasında oluşan ilişkiler	9
Şekil 2.4. Königsberg'in 2 adası ve 7 köprüsü	10
Şekil 2.5. Düğümler, kenarlar ve kenarların düğüm kümeleri	11
Şekil 2.6. Yönlü ve yönsüz çizgeler	12
Şekil 2.7. Ağ analiz yazılımları ağı	14
Şekil 2.8. B düğümü A ile diğer düğümler arasında	16
Şekil 2.9. B düğümü kümelenme katsayısı	17
Şekil 3.1. Sosyal ağın çizge yapısı	20
Şekil 3.2. Sosyal ağın zamansal değişimi	23
Şekil 3.3. Amerika'daki bir okuldaki arkadaşlık ağı	25
Şekil 3.4. İki mod ilişki matrisi	26
Şekil 3.5. İki moddan tek moda dönüştürme	27
Şekil 3.6. Komşuluk matrisi	27
Şekil 3.7. Bağlantı tahminine farklı yaklaşımlar	28
Şekil 3.8. Ağ yapısındaki ilişkilerin zamansal değişimi	29
Şekil 4.1. Hastalık ilaç ağı	38
Şekil 4.2. Hastalık ağı	39
Şekil 4.3. İlaç ağı	40
Şekil 4.4. İlaçların birlikte yazılma bağlantı tahmini grafiği	42
Şekil 4.5. Ağrı kesici ilaçların bağlantı tahmini sonuçları (Kaynak Tahsisli)	44
Şekil 4.6. Ağrı kesici ilaçların bağlantı tahmini sonuçları (Adamic Adar)	45
Şekil 4.7. Ağrı kesici ilaçların bağlantı tahmini sonuçları (Jaccard)	47

TABLÖLAR LİSTESİ

Sayfa No:

Tablo 2.1. Farklı bilim dallarında ağların farklı kelimelerle ifadesi	10
Tablo 3.1. Sosyal ağ araştırmalarında karşılaşılan bağ türleri ve veriseti özellikleri.....	25
Tablo 3.2. Yerel benzerlik algoritmaları karşılaştırması	33
Tablo 4.1. Veritabanında kullanılan hastalıklar ve ilaçlar	37
Tablo 4.2. Hastalık bağlantı tahmini değerleri	40
Tablo 4.3. İlaç bağlantı tahmini değerleri	41
Tablo 4.4. Ağrı kesici ilaçların bağlantı tahmini değerleri (Kaynak Tahsisli).....	43
Tablo 4.5. Ağrı kesici ilaçların bağlantı tahmini değerleri (Adamic Adar).....	45
Tablo 4.6. Ağrı kesici ilaçların bağlantı tahmini değerleri (Jaccard).....	46

KISALTMALAR LİSTESİ

SNA	:Social Network Analysis
ABD	:Amerika Birleşik Devletleri
LHN2	:Leicht Holme Newman 2
RWR	:Random Walk with Restart
MFI	:Matrix Forest Index
LRW	:Lokal Random Walk
SRW	:Superposed Random Walk

1.GİRİŞ

Bilgi güçtür. Bilgi çağında değer yaratmanın yolu fiziksel üretimden ziyade, bilgi kaynaklarını etkin ve yararlı kullanmaktan geçer. Bilgiye sahip olmanın ve onu kullanmanın önemli olduğu günümüzde güçler dengesi bilgi üzerine yoğunlaşmıştır. Bilgiyi elde edip kullanan toplumlar daha hızlı gelişir. Bilgi sayesinde her alanda gelişme gösterir ve buna bağlı olarak gelirlerini ve refah düzeylerini artırdıkları gibi, dünyada güçlü konuma gelirler.

Çeşitli sebeplerden dolayı bilişim teknolojisi hayatımızın her alanında herkes tarafından kullanılmaya başlandı ve yaygınlığı çok süratli bir şekilde artmaya devam etmektedir. Bu sebeple veri miktarları inanılmaz şekilde artmaktadır.

Günümüzde gelişen bilgi teknolojileri sayesinde her geçen gün daha çok veri sayısal olarak toplanıp, saklanıp ve hepsinden önemlisi kullanılmaktadır. Veri bilgiye dönüştürülüp, kullanılıncaya kadar değer ifade etmez. Değerli olan verilerin irdelenip, bilgiye dönüştürülmesi ve karar için kullanılabilmesidir.

Yine aynı zamanda bu verilerin çoğu insanların birbirleriyle olan ilişkileri üzerine oluşmaktadır. Yani insanların etkileşimi ile bu veriler ortaya çıkmaktadır. Bu yüzden bu yapılar çeşitli durumlarda yapılan işlemlerin sonucudur. Bu verilerde etkileşim ve ilişki oluşumu olduğu için sosyal ağ adını verdiğimiz bir yapı ortaya çıkmaktadır. Çevremizde sosyal ağ yapısı olarak adlandırabileceğimiz birçok yapı vardır.

Bilgisayar ortamında insanların birbirleriyle ilişkiler oluşturmalarını sağlayan yeni nesil uygulama sosyal ağlardır. İnsanların birbirleri ile etkileşime geçtiği bu ortamlar bir yol haritası gibidir. Bu ağlar analiz edildiğinde işimize yarayacak çok değerli bilgiler elde edilebilir. Bu ağlara örnek olarak Facebook, Flickr, Friendfeed, Myspace, Twitter, Youtube, Blogger, LinkedIn, Google+ gibi siteler ve etkileşim üzerine yapılan yazılım uygulamaları verilebilir. Bilgi çağında bu ağlardaki bireyler arasında oluşan ilişkiler sayesinde bireylerin sosyal davranışlarını inceleyebilir, bu ilişkiler hakkında çeşitli değerlendirmelerde bulunabilir ve bunlardan işe yarar bilgiler elde edilebilir. Bu bağlantılardan gelecekteki durumlar içinde çeşitli tahminlerde bulunulabilir.

Bağlantı tahmini konusu makine öğrenmesi alanında son yıllarda üzerinde çalışmaların artarak devam ettiği bir alandır. Günümüzde bibliyografik alan [1], moleküler biyoloji [2], adli soruşturmalar [3], tavsiye amaçlı sistemler [4], tıp [5] vb. alanlarda bağlantı tahmini yapısı kullanılmaktadır. Dinamik bir yapıya sahip ağlarda bağlantı

tahmininde bulunmak zor bir konudur. Ağ yapısında bulunan çeşitli verilerden yeni veriler elde etme ve geleceğe dönük tahminler üretmek amaçlanmaktadır. Bu durumda ağın gelecekteki yapısı ve bireylerin oluşturabileceği yeni ilişkilerin veya vazgeçeceği ilişkilerin tahmini yapılmaya çalışılmaktadır.

Bu tez kapsamında sosyal ağlarda bağlantı tahmininin tanımı, özellikleri, bu konuda yapılmış çalışmalar incelenip, hastalık ilaç verileri kullanılarak sosyal ağ analizi teknikleri ile bağlantı tahmini uygulaması gerçekleştirildi. Çalışmanın bölüm organizasyonları aşağıdaki şekilde düzenlendi.

İkinci bölümde öncelikle uygulamada sosyal ağ yapıları kullanılacağından sosyal ağ yapılarından bahsedildi. Sosyal ağ analizi için kullanılan çizge yapılarına değinildi. Daha sonra analiz için kullanılan çeşitli yazılımlar incelenerek, ağ analiz aşamaları, ölçütleri anlatıldı.

Üçüncü bölümde bireylerin birbiriyle ilişki bağları kurmaları sayesinde oluşan yapı üzerinden yola çıkılarak ilişkiye yönelik analiz yapmayı sağlayan bağlantı tahmini adı verilen yöntemler incelendi.

Dördüncü bölümde yapay oluşturulan veritabanı üzerinde çeşitli bağlantı tahmini yöntemleri kullanılarak uygulamalar yapıldı. Bu uygulama sonucunda elde edilen veriler değerlendirildi.

Beşinci bölümde ise yapılan çalışma değerlendirilip sonuçlandırıldı.

2.SOSYAL AĞLAR VE ANALİZİ

Bilgisayar sistemleri her geçen gün hem daha ucuzlamakta, hem de güçleri artmaktadır. Artık bilgisayarlar daha büyük miktardaki verileri daha hızlı bir şekilde işleyebilmektedir. Aynı zamanda ağ haberleşmesinin gelişmesiyle farklı bilgisayarlar dağıtık olarak aynı işler için bir arada kullanılabilmektedir. Çeşitli sebeplerden dolayı bilişim teknolojisi hayatımızın her alanında herkes tarafından kullanılıp ve yaygınlığı çok süratli bir şekilde artmaya devam etmektedir. Bu sebeple veri miktarları inanılmaz şekilde artmaktadır [6].

Yaşamış olduğumuz bu zaman diliminde her alışverişte, her bankacılık işleminde, her türlü kamusal alandaki işlemlerde kayıt edilen veriler bulunmaktadır. Ayrıca işletmelerin ve devletin kendi bünyelerinde yapılan işlemler sonucunda da sakladığı veriler, görüntü ve ses cihazlarından elde edilen çoklu ortam verileri gibi birçok veri sürekli olarak depolanmakta ve depolanan bu veriler çok hızlı boyutlarda artmaktadır. Fakat bu veriler istenildiği şekilde değerlendirilememekte, hızla büyüyen bilgi yığınları şekline dönüşmektedir.

Her geçen gün işletmeler ve devlet kurumları veri tabanı sistemine daha fazla yatırım yapmakta ve daha fazla veriyi bu sistemlerde depolamaktadır. Fakat bu boyuttaki veri çoğu işletme veya kurumda anlamlı ve verimli bir şekilde işlenememektedir [7].

Günümüzde bilişim sistemlerinin hayatın hemen her alanında aktif bir rol oynar hale gelmesi ile birlikte; veri, enformasyon, bilgi vb. birçok kavrama aşına olmuş durumdayız. Özellikle kişisel bilgisayarların günlük yaşamın herhangi bir kademesinde karşımıza çıkıyor olması, bilgisayar dünyasının jargonunu dilimize güncel olarak sokmaktadır. Çevremizdeki birçok insan “data”lardan “information”lardan söz etmektedir. Türkçeleştirilmiş haliyle veri ve bilgi kelimeleri içimizde yaşayan birer birey halini almış durumdadır. Günlük kullanımda sıkça telaffuz edilen bu kavramlar, salt birer kelime olarak kullanıldıklarında anlam daralmasına hatta anlam kaymasına maruz kalmaktadırlar. Oysa kavram olarak ele alındıklarında oldukça önemli bir yere sahip olan bu hece toplulukları, bilişim dünyasının hiç şüphesiz yapı taşı konumundadırlar. Bu sebeptir ki; bu kavramları doğru adreslemek gerekmektedir.

Bilginin bu denli değerli olması, bilişim teknolojilerinin gelişmesine ön ayak olmuştur. Bilgisayarların bilgi yönetiminde ve üretimde faal olarak yer almaya başlaması

kaçınılmaz bir durum haline gelmiştir. Günümüze bakıldığında bir bilgi patlaması söz konusudur. Çevremizin verilerle dolu olması peşi sıra enformasyon ve bilgiyi beraberinde getirmektedir. İnternet gibi etkili bir iletişim ortamının varlığı bu durumu körüklemektedir. Makro düzeyde bakıldığında hemen hemen herkes bu veri dağına bir katkıda bulunmakta ve bundan yararlanmaktadır. Ancak bu yanında bazı sorunlar da getirmektedir. Bu kadar çok veri arasından gereken bilgiyi çıkartabilmek gerekmektedir. Bu aşamada yeni bir kavram karşımıza çıkmaktadır: Veri Madenciliği [8].

Veri madenciliği büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanarak aranmasıdır. Geleceğin, en azından yakın geleceğin, geçmişten çok fazla farklı olmayacağını varsayarsak geçmiş veriden çıkarılmış olan kurallar gelecekte de geçerli olacak ve ilerisi için doğru tahmin yapmamızı sağlayacaktır [6]. Veri madenciliği yöntemleri kullanarak gelecekle ilgili tahminler ve analizler yapmanın en çok kullanıldığı alanlardan birisi de sosyal ağlardır.

Sosyal ağların önemi dünyanın küçüklüğü ile ilgili bir olgudur. 1970’li yıllara kadar dünyanın çok büyük olduğu ve insanların birbirini tanımadığı düşüncesi vardı. Küçük dünya fikrini ortaya atan Sosyolog Stanley Milgram bu düşüncüyü değiştirmiştir [9]. 1967’de bir posta deneyiyle dünyanın küçük olduğunu ispatlamıştır. Harvard’dan tesadüfi olarak Omaha, Nebraska’dan seçtiği yaklaşık 300 kişiye mektuplar yollayarak, onlardan bu mektubu Boston’daki hedef kişiye sadece kişisel kontaklarını kullanarak iletmesini istemişti. Milgram, mektup yolladığı kişilere, ulaşmaları gereken kişinin ismini, yerleşimini, mesleğini vermişti. Nebraska’lı 300 kişi Boston’dakini tanımasa da, onu tanıma ihtimali olan kişiler aracılığı ile hedefe yaklaştı. Aile bireylerinden biri, iş arkadaşı, okul arkadaşı vb. yardımı ile mektuplar seyahatlerine başladı. Milgram, hedefe 60 zincirin ulaştığını gördü. Bu zincirlerde de ortalama 6 basamak olduğu için, sonuç daha sonraları tiyatro oyunlarına, filmlere isim olan *Altı Derece Uzak* olarak literatüre girdi. 2003’te Columbia Üniversitesi’nden bir grup bu deneyi e-posta ile 166 ülkeden 60 bin kişiyle tekrarlayarak Altı Derece Uzak’ın geçerliliğini bir kez daha kanıtladı. Deneyde ortaya çıkan altı dereceli sınıflandırmada, bir kişi tanımadığı birisine en fazla 5 kişiyi kullanarak ulaşabilmektedir [10]. Her ne kadar Milgram bu terimi kullanmasa da bu ve benzeri deneyler “*Six Degrees of Separation*” yani “*Ayrılığın Altı Derecesi*” kavramı ile ilişkilendirildi ve çevrimiçi sosyal ağların gelişimine de zemin hazırladı. Zaten ilk kurulan sosyal ağın adı da *sixdegrees.com*’dur.

Günümüzde iletişim teknolojilerinin gelişerek sosyal yaşamı derinden etkilemesi ile insanların birbirleriyle olan ilişkilerini sanal ortam üzerinde paylaşp, yönetmelerine dayanan yapıları yani sosyal ağ adını verdiğimiz yapıların ortaya çıkmasını sağladı. Günümüzde tüm dünyada beklentilerin üstünde bir hızla genişleyen ve neredeyse bilgisayar kullanan her insanı etkisi altına alan sosyal ağlar, sanal ortamda insanların birbirleriyle olan ilişkilerini kontrol edebilmesini sağlayan yeni nesil web uygulamalarıdır. Sosyal ağ siteleri, sosyal bir çevre oluşturmak amacıyla kurulan, büyük kitlelerin birbirleriyle yoğun olarak iletişime ve etkileşime geçebildikleri elektronik ortamlardır. Özmen [11], sosyal ağ sitelerini kişilerin, kendilerine ait bir alan oluşturarak sistemdeki diğer kullanıcılarla bir araya gelmelerine, dijital ortamda paylaşılacak her türlü materyali (video, fotoğraf, haber, vb.) birbirleriyle paylaşmalarına ve etkileşim içine girmelerine imkân sağlayan sanal ortamlar olarak tanımlamıştır. Boyd ve Ellison [12], sosyal ağları bireylerin sınırlı bir sistem içerisinde halka açık veya yarı açık bir profil oluşturmalarına, bağlantılı oldukları diğer kullanıcıların bir listesini beyan etmelerine, kendilerinin ve sistemdeki diğer kullanıcıların bağlantı listelerini görüntülemelerine ve geçiş yapmalarına izin veren web tabanlı servisler olarak tanımlamıştır. Carminati ve arkadaşları [13] web tabanlı sosyal ağları; web kullanıcılarının kaynak paylaşımına ve iş, eğlence, flört gibi amaçlarla diğer kullanıcılarla arkadaşlık kurmalarına olanak sağlayan çevrimiçi topluluklar olarak tanımlamıştır.

Bugün yerli, yabancı yüzlerce sosyal ağ sitesi bulunmakta ve bu siteler işleyişleri bakımından genel olarak iki grup altında incelenmektedir [14].

Birinci grup; Bireylerin ağırlıklı olarak “kişisel bağları” bulunan insanlarla etkileşime geçtikleri kullanıcı tabanlı sosyal ağ siteleridir. Facebook, Myspace, Orkut gibi, özel bir konu başlığı olmaksızın, genel kullanıcılara hitap ederler.

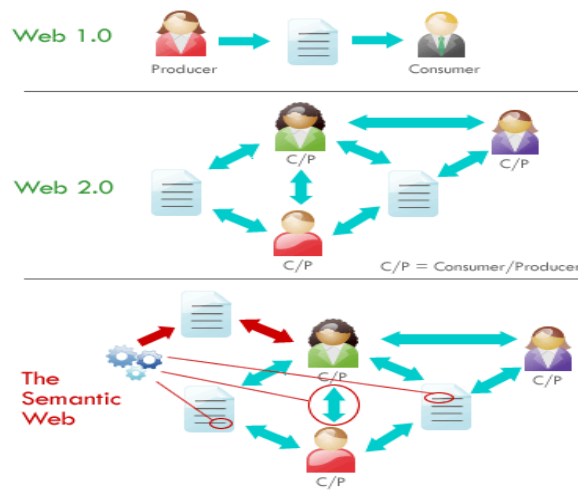
İkinci grup; İnsanları belirli bir konu, hobi veya düşünce etrafında toplayan grup tabanlı sosyal ağ siteleridir. Üye sayıları genel olarak kullanıcı tabanlı sosyal ağlardan daha azdır ve kayıt kuralları daha katıdır. Belli bir ilgi veya ilişki esas alınarak oluştururlar. LinkedIn, Academia ve Ravelry bu grup sosyal ağlara örnek gösterilebilir [15].

İnternet ortamında tanıdıklarıyla bağlantı kuran insanlar kendilerine sanal bir dünya yaratmaktadırlar. Kişilerin tanıdığı veya tanımadığı kişilerle etkileşime geçebildiği, içerik üretebildiği, kendi dünyasını kurduğu ortamlardır. İlk başlarda web teknolojisi bu dinamikliğe sahip değildi. Web teknolojisi gelişimi aşamaları Şekil 2.1’de görüldüğü gibidir. Bu aşamalardan özellikle Web 2.0’den sonra hızlı bir büyümeye giren sosyal ağlar

insanların yaşamını etkilemeye başladı. Kullanım oranlarının her geçen gün giderek artması, özellikle internet teknolojilerinin hızla gelişmesi ve dünya çapında yaygınlaşması ile birlikte, insanlar kendilerini çok daha rahat ifade edebildikleri sanal dünya üzerinde daha fazla vakit geçirmeye başladı. Günümüz dünyasında gerçek hayatta insanların sosyalliğinin azalması, fitri yapısında sosyallik olan insanların kendilerine rahatça ifade edebildiği, daha rahat ilişkiler kurabildiği bu ortamları ortaya çıkardı. Sınırların ortadan kalktığı bu dünyada, farklı coğrafyalardan, farklı kültürlerden bireyler birbirleriyle kolay ve hızlı bir şekilde iletişim kurabilmektedirler.

Büyümesini Web 2.0'ye bağladığımız sosyal ağların aslında internetin doğuşundan bu yana kullanıldığı gerçeğini değiştirmez. Örneğin, kendi aralarında e-posta yoluyla haberleşen kullanıcıların oluşturduğu bir sistem içinde barındıracağı çizgesel veri yapısı nedeniyle sosyal bir ilişkiye dayalı bir ağ gibi düşünülebilir. Bunun dışında anlık mesajlaşma uygulamalarındaki (Skype, MSN Messenger vb.) görüşme listeleri de veri yapılarından dolayı sosyal ağ olarak görülebilir.

Web 2.0 ile beraber internette pek çok alanda gerek programlama ve tasarım gerekse web sitelerinin yapısı olarak büyük değişime gidildi. Dinamik bir yapıya kavuşan internet sitelerinin içeriğinin kullanıcılar tarafından oluşturulması, yapısal olarak internet yaşamını değiştirdi. İnternet siteleri artık sadece ortamı oluşturup, gerisini kullanıcıya bıraktı. Facebook, Flickr, Friendfeed, Myspace, Twitter, Youtube, Blogger, LinkedIn, Google+ gibi siteler bu temel üzerine kuruldu.



Şekil 2.1. Yeni nesil web

İnternet kullanıcılarının birbirleriyle tanışması, irtibata geçmesi, içerik paylaşımında bulunması, tartışma ortamı oluşturması ve ortak ilgi alanlarındaki kişilerin bir araya gelebileceği gruplar oluşturması amacıyla oluşturulan internet siteleri sosyal ağlar olarak tanımlanmaktadır. Ülkemizde oldukça popüler olan Facebook ve Twitter bunun en iyi örnekleridir. Sosyal ağın ne olduğunu bilmeyenler bile Facebook ve Twitter'ı bilir. Ayrıca kişilerin iş ağını genişletebileceği LinkedIn iş dünyasının en önemli sosyal ağlarından [16]. Şekil 2.2'de sınıflandırılmış farklı birçok sosyal ağın mevcut olduğu görülebilmektedir.



Şekil 2.2. Sosyal ağ sınıfları

Sosyal ağların herkes tarafından kullanıldığı bir gerçek, 76 milyonluk ülkemizde 26 milyonluk, 7 milyarlık dünyamızda ise 650 milyonluk facebook kullanıcısı varken veri madenciliği için en önemli alanlardan biri sosyal ağlardır [17]. Sosyal ağ analizinde en

önemli faktör güçlü bir veri kaynağıdır. Bu veri kaynağı da milyonlarca üyeye sahip platformlarda fazlasıyla bulunmaktadır. Bu platformlar, kişilerin internete bağlanma alışkanlıklarını ve diğer kişilerle kurdukları bağları ortaya koyan ham veri için bulunmaz bir kaynaktır. Zaten sosyal ağlar bu kadar değerli kılan da sahip oldukları bu veri kaynaklarıdır. Sosyal ağlar, milyonlarca değişik insanın davranışlarına ilişkin oldukça fazla bilgi içermektedir. Bu davranışların, etkileşimlerin analiz edilmesi üzerinden çeşitli tahminlere varılabilir. Bu ağların analizi bize birçok yararlı bilgi verebilir.

Bu bakış açısından yola çıkılarak ağ yapılarının içerisindeki varlıklar arası ilişkilerin çeşitli bilimsel metotlar aracılığı ile detaylı olarak incelenmesi sonucu elde edilen verilerden anlamlı sonuçlar türetilmesi işi ise “Sosyal Ağ Analizi” olarak tanımlanmaktadır. Bireyler arası ilişkilerin sayısallaştırılıp bilimsel hale getirilmesi demek olan sosyal ağ analizi, önemli olaylar karşısında çeşitli organizasyonların, ya da bu organizasyonların oluşturduğu ağların da ilişkilerini rakama dökmek için kullanılmaktadır. Bilgisayar programlarına girilen verilere göre alınacak olan çıktının niteliği de değişmekte ve bu esneklik organizasyonel verimliliği test etmek için kullanılabilecek yeni bir olanak sağlamaktadır.

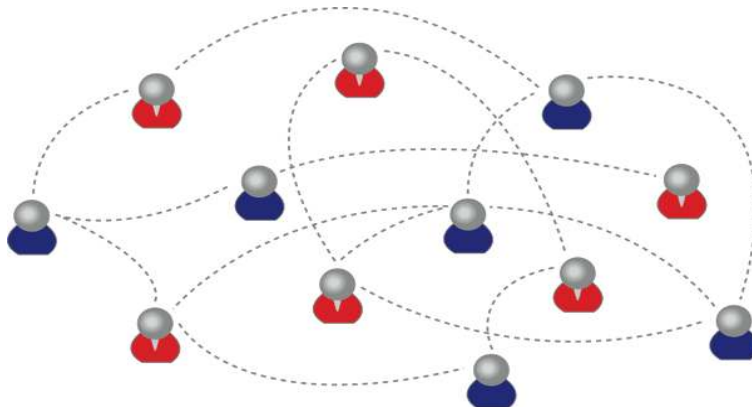
Sosyal ağ analizi yeni bir kavram değildir ve 19. yy. sonlarından itibaren sosyoloji biliminin öncülüğünde gelişim göstermiştir. Teknolojide yaşanan büyük gelişmelerin, bireylerin ve toplulukların etkileşimini çok farklı boyutlara taşıması ile daha da gelişmiştir.

Günümüzde iletişim teknolojilerinde ve özellikle internet ortamında yaşanan gelişmeler neticesinde bireyler arasındaki etkileşim daha hızlı ve karmaşık bir hal almış durumdadır. İki kişi arasındaki iletişim zaman ve mekândan bağımsız bir hale gelmiştir. İnternet üzerindeki sosyal paylaşım siteleri, bloglar, forum ortamları, mesajlaşma yazılımları ve daha bunun gibi pek çok ortam insanlar arasındaki ilişkilerin farklı bir boyut kazanmasına neden olmuştur. Sosyal ağ analizi günümüzde pek çok alanda kullanılmakta olup bunların başında birey ve sosyal grup yapılarının ve davranışlarının incelenmesi, elektronik ticaret ve çevrimiçi reklamcılık, fiziki yapıların analizi ve büyük veri kümelerinin analizi gelmektedir. Eski Irak Kralı Saddam Hüseyin’in yakalanmasında bireysel ağ ilişkilerini etkili bir biçimde ortaya koyan UCINET isimli programın Amerikan ordusu tarafından nasıl kullanıldığı, yine başka bir terör ağının Sosyal Ağ Analizi (SNA) yardımıyla nasıl ortaya çıkarıldığı bilinmektedir.

Sayısal ortamlarda sağlanan iletişimin yarattığı imkânlardan bir diğeri ise oluşan iletişim verilerinin somut bir biçimde ölçülebilir olmasıdır. Çeşitli ortamlarda saklanan

veriler, geliştirilen teknikler sayesinde detaylı olarak analiz edilmekte ve bunun sonucunda bir takım değerli bilgilere ulaşılmaktadır. Ancak, sayısal ortamlar üzerinde oluşan sosyal ağ yapıları içerisindeki aktörlerin birbirleri ile olan ilişkileri, etkileşimleri ve bilgi paylaşımları sonucu ortaya çok büyük ölçekte bir veri kümesi çıkmaktadır. Bu veri yığınının işe yarar verilerin ve ileriye yönelik kuvvetli tahminlerin ortaya çıkarılmasında “Veri Madenciliği Teknikleri” büyük fayda sağlamaktadır [18].

Sosyal ağlar insanlık tarihi kadar eski ilişkilerdir. Şekil 2.3’te görüldüğü gibi insanlar arasındaki politik, resmi-gayri resmi, ailevi, coğrafi ya da herhangi başka bir şekilde ilişkiler sosyal ağları oluşturur. Bu ağları analiz etmek için kullanılan bilgisayar teknolojilerinin artan miktardaki yazılımı ve kullanımı, sosyal ağ analizi yöntemini akademik ve praktisyen sahalar için erişilebilir konuma getirmektedir. Halen bu alanda geliştirilmiş birçok bilgisayar programı olması ve bir yenisinin her gün literatüre eklenmesi bu alanın gelecekte ne kadar gelişeceğini de göstergesidir. Genelde (özellikle Batı Avrupa’da) Sosyal Ağ Analizi, ekonomik ilişkileri ortaya koymak amacıyla kar amacı güden organizasyonların bağlantılarını işlemek üzere kullanılırken, ABD’de her türlü ikili ya da daha çoklu ilişkiyi ortaya çıkarmak için kullanılmaktadır. Özellikle sosyal ilişki kurma amacıyla kurulmuş internet sitelerinde bireylerin diğerleriyle kurdukları kontakların bilimsel dilde anlaşılabilmesi için Sosyal Ağ Analizi yoğun bir şekilde kullanılmaktadır. Yine bu amaçla bu sitelerde (Facebook, MySpace, Linked-in vs.) sosyal ağ grupları kurulmakta ve bilginin bireyler arasında ne yönde taşındığı konusunda araştırmalar ortaya konmaktadır [19].



Şekil 2.3. Düğümler arasında oluşan ilişkiler

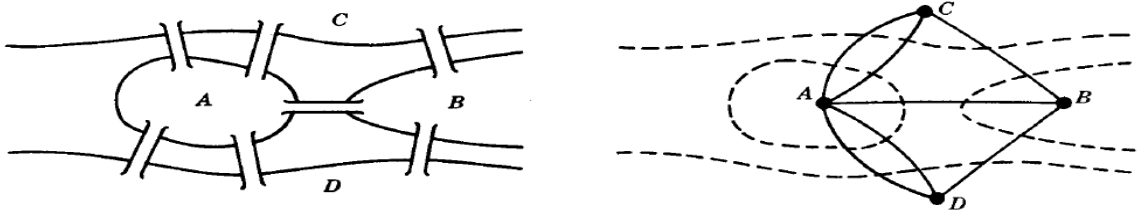
2.1 Çizge Teorisi

Matematik ve bilgisayar bilimlerinde yoğun olarak kullanılan çizge teorisinin uygulamaları modern hayatın karmaşık ve geniş kapsamlı birçok probleminin çözümü için kullanılmaktadır. Çizge teorisi problemleri tanımlama ve yapısal olarak ilişkileri belirlemekte faydalıdır. Basitçe bir çizge düğüm olarak adlandırılan noktalar ve her biri bu noktaları veya sadece noktanın kendisini birleştiren ve kenar olarak adlandırılan çizgilerdir. Örnek olarak şehirleri düğüm ve onları bağlayan yolları kenar olarak gösteren yol haritaları verilebilir. Tablo 2.1’de gösterildiği gibi farklı bilim dallarında tanımlamalar değişebilmektedir. Bir çizgeyi tanımlamak için öncelikle düğümlerin ve kenarların kümesini tanımlamak gerekir. Daha sonra hangi kenarların hangi düğümlere bağlandığı gösterilmelidir. Bir kenarın her iki ucunda da düğüm olacak şekilde tanımlanmalıdır.

Tablo 2.1. Farklı bilim dallarında ağların farklı kelimelerle ifadesi [20].

	Mühendislik ve Bilgisayar Bilimleri	Matematik	Fizik	Sosyal Bilimler
“nokta”	düğüm	vertex	yer	aktör
“çizgi”	bağlantı	kenar	bağ	beraberlik
“ağ”	ağ	graf	ağ	ağ

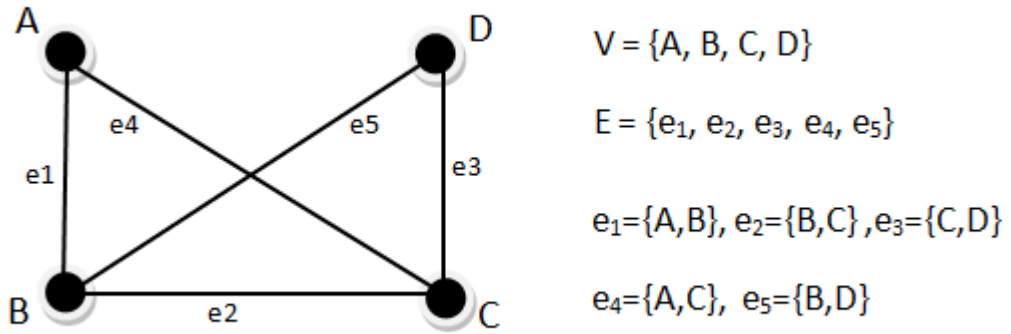
Aşağıdaki Şekil 2.4’de görüldüğü gibi çizge teorisinin çıkış noktası 18. yüzyıl Doğu Prusya kasabasıdır. Königsberg’in 2 adası ve 7 köprüsü vardır. Königsberg halkı ünlü matematikçi Euler’e, bir kişinin herhangi bir yerden başlayıp herhangi bir yerde durarak ve her köprüyü bir ve en fazla bir kez geçerek bir gezinti yapıp yapamayacağını sormuşlardır. Euler problemi şekildeki gibi çizge ile çözmeye çalışmıştır.



Şekil 2.4. Königsberg’in 2 adası ve 7 köprüsü

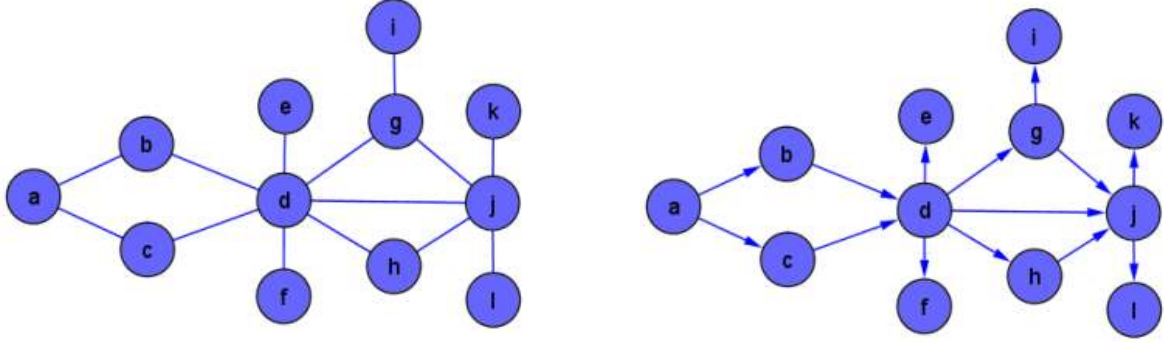
Sosyal ağlar, bilgisayar ortamında çizge olarak temsil edilebilirler. Genel çizge gösterimi $G = (V, E)$ şeklindedir. Bu gösterimdeki V sonlu düğüm kümesini ve E sonlu kenar kümesini belirtir. Her bir kenar kendisi ile ilişkilendirmiş bir ya da iki adet düğüm içerir ve bu düğümleri birbirine bağlar. Sosyal ağlardaki kullanıcılar (insan, grup, kurum vb.) çizgede düğüm olarak temsil edilirken, sosyal ilişkiler (arkadaşlık, profesyonel iş ilişkisi vb.) çizgede kenar olarak temsil edilirler. Bazı çizgelerde iki düğüm arasında birden fazla kenar bulunabilir. Eğer bir $E = \{U, V\}$ kenarı varsa U ve V düğümünün komşu olduğu söylenir. Böylesi bir durumda U ve V , E 'nin uç noktaları olarak adlandırılır ve E 'nin U ve V 'yi bağladığı söylenir.

Çizgeler, düzlemsel diyagramlarla gösterilir. Şekil 2.5'de gösterildiği gibi V kümesindeki her v düğümü bir nokta (ya da küçük çember) ile temsil edilir ve her $E = \{v_1, v_2\}$ kenarı, v_1 ve v_2 uç noktalarını bağlayan bir çizgi ile gösterilir.



Şekil 2.5. Düğümler, kenarlar ve kenarların düğüm kümeleri

Aynı uç noktalarını bağlayan çoklu kenarla veya uç noktaları tek ve aynı düğüm olan döngüler barındıran çizgelere çoklu çizgeler denir. Çizgelerin resmi tanımları ne çoklu kenarlara ne de döngülere izin verir. Bazen, çizge terimine yüklenen anlam çoklu çizgeleri de içerir ve çoklu kenar ve döngü içermeyen çizgeler için basit çizge terimi kullanılır. Bir kenar bir düğümü kendisine de bağlayabilir. Bu tür döngüler çizge içerisinde yer alabilir. Şekil 2.6'da gösterildiği gibi kenarlar yönlü de olabilir. Yönlü çizgeler adı verilen bu çizge yapılarında kenarların başlangıç ve bitiş noktaları bellidir. Yönlü ve yönsüz kenarların birlikte bulunduğu karışık çizgelerde mevcuttur. Bu tür çizge özellikleri daha da sıralanabilir. Günlük hayatta pek çok farklı alanda kullanılan çizge modelleri sosyal ağları temsil etmek için kullanılabilir [21].



Şekil 2.6. Yönlü ve yönsüz çizgeler

2.2. Sosyal Ağ Analizi Yazılımları

Bu alandaki gelişmelerle beraber ağ analizi için birçok yazılım geliştirilmiştir. Başlıca sosyal ağ analizi araçları Ucinet, NetMiner, Pajek, ORA, StatNet, SocNet-V, InFlow ve Keyhubs'tır. İlk iki yazılım ABD'de yoğun bir şekilde kullanılmaktadır. Türkçe bir sosyal ağ analiz yazılımı maalesef henüz piyasada bulunmamaktadır.

Bu yazılımlar genelde ağ aktörleriyle ilgili belli bir içerikte bulunan karşılıklı (iki-yönlü) ya da direk (tek-yönlü) ilişkileri bir matris tablosunda inceleyerek ilişkilerin yönü ve konumu konusunda bilgi sağlar. Dolayısıyla içerik, ağ aktörleri ve incelenen ilişki cinsi çok önemlidir. İçerik; gazete haberleri, resmi raporlar, bireysel anketler ya da ilişki sorgulayan diğer enstrümanlar olabilir. İçeriğin objektif olması sonuçların da objektif olmasını sağlar.

Sosyal ağ analiz yazılımları, içerikten oluşturulan ilişki matrisinden aldığı bilgiyi grafikler veya artan-azalan veri tabloları yoluyla kullanıcıya geri bildirir. Bunlara kısaca değinmek gerekirse derece merkeziliği bir üyenin (düğüm) diğerleri arasında ilişki olsun ya da olmasın bunlarla ilişkili olmayı, arasındalık merkeziliği iki ya da daha farklı üye grubu arasındaki kilit üyeliği, yakınlık merkeziliği kilit konumdaki üyelere daha yakın olmayı betimler. Literatürde, bir üyenin ağın kilit üyesi olup olmamasını derece merkeziliği veya arasındalık merkeziliği belirler [19].

Aşağıda bazı ağ analiz programlarından bahsedilmektedir;

Ucinet: Kapsamlı bir sosyal ağ veri analiz paketidir. En sık kullanılan yazılımlardan biridir. Steve Borgatti ve arkadaşları tarafından geliştirilmiştir. İki mod ve tek mod veri üzerinde işlem yapılabilir. Farklı biçimlendirilmiş metin dosyalarını ve Excel dosyalarını okuyabilmekte ve yazabilmektedir. 32767 düğüme kadar veri üzerinde işlem yapabilmektedir. Merkezilik ölçütleri, alt grup tanımlama, rol analizi, temel çizge teorisi işlemleri ve permütasyon tabanlı istatistiksel analiz işlemleri gibi yöntemler kullanılarak sosyal ağ analizi yapılabilir. Ucinet güçlü matris analiz yöntemleri barındırmaktadır. Ağ görselleştirmesi için NetDraw aracı dâhili yüklenmektedir. Veri kaydedilirken Pajek ve Mage programlarında çalıştırılacak şekilde kaydedilebilmektedir. KrackPlot, NEGOPY ve Pajek uygulamalarının desteklediği formatlarda veri yükleyebilme özelliği de bulunmaktadır.

Pajek: Büyük miktarda veri setlerini işleyebilmek için tasarlanmış bir ağ analiz ve görselleştirme uygulamasıdır. Ücretsiz bir yazılımdır. Ağı alt ağlara dönüştürme, görselleştirme ve analizler yapılabilir. Bir milyondan fazla düğüme sahip ağları analiz edebilmektedir. Matris tabanlı çalışmaktadır. Ucinet DL formatında veri girişi vardır. Windows ortamında çalışmaktadır. Vladimir Batagelj ve Andrej Mrvar tarafından geliştirilmiştir. Bazı yordamları da Matjaž Zaveršnik tarafından yazılmıştır.

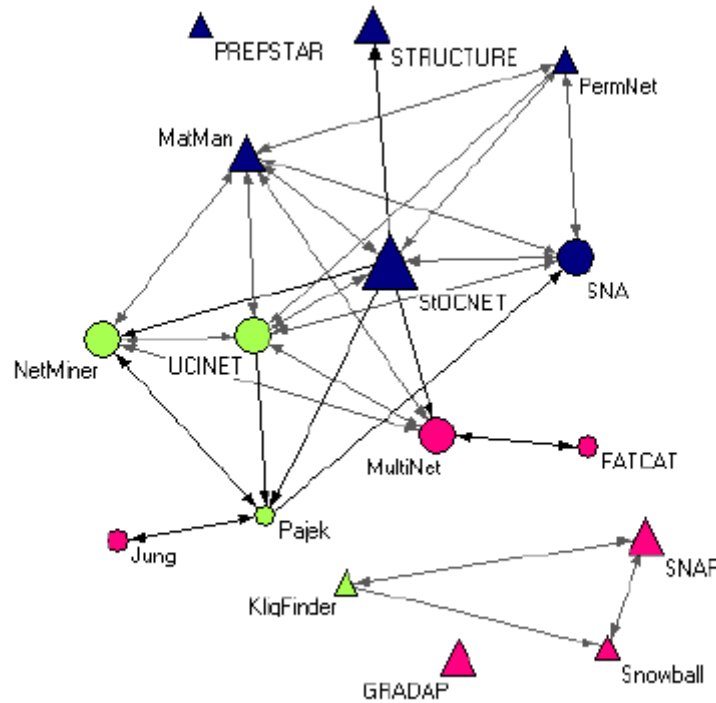
Igraph: C/C++ dilinde hazırlanmış bir kütüphanedir. Yönlü ve yönsüz ağların analizi için literatürde kullanılan algoritma ve modelleri içeren birçok fonksiyon barındırmaktadır. Ayrıca çizge teorisi problemlerine de çözümler sunmaktadır. 2D ve 3D boyutlarında ağ görselleştirmesi yapılabilir. Kendi algoritmalarını farklı dillerde gerçekleştirme imkânı sağlamaktadır. GraphML, GML, Pajek ve başka dosya uzantılarını desteklemektedir. Açık kaynak kodludur. Farklı işletim sistemlerinde çalışan sürümleri mevcuttur.

Multinet: Dağınık ve geniş ağların analizi için kullanılan bir yazılımdır. Matris tabanlı çalışmaktadır. Multinet'in bazı özelliklerini barındıran FATCAT gibi alt programları da vardır. ASCII ve CSV uzantılı dosyaları desteklemektedir. İstatistiksel analiz yöntemleriyle analizler yapmaktadır. Linklerin yönü ve ağırlığı ve düğümlerin özellikleri spektral analiz için grafiklerde gösterilebilmektedir. Değişkenler üzerinde birçok çeşit dönüşüm ve kayıt yapılabilir. Verilerdeki hatalar tespit edilebilmektedir.

Statnet: Ağ görselleştirme, simülasyon ve analiz yapabilen bir yazılımdır. Yazılım C dili ve istatistiksel bir dil olan R ile geliştirilmiştir. Yönlü, yönsüz, tek mod ve çift mod ağların analizinde kullanılabilmektedir. R dilinin komut ara yüzünü kullanmaktadır. NETWORK, SNA ve ERGM yazılımlarını içinde barındıran bir pakettir. Ağ analizi için kullanılan kapsamlı bir yazılımdır.

Netminer: Büyük ağların analizi için kullanılan diğer bir yazılımdır. Veri dönüştürme, ağ analizi, istatistiksel işlemler, görselleştirme, grafik tabloları oluşturulabilmektedir. Python dilinde geliştirmeler yapılabilmektedir. Ticari amaçlı geliştirilmiş bir program olmasına rağmen akademik çalışmalar içinde kullanıma izin verilmektedir. Metin dosyaları, Ucinet, Pajek, StOCNET, Excel dosyalarından veri girişi ve çıkışı yapılabilmektedir. GML dosyalarından da veri girişi yapılabilmektedir. Çok özellikli bir yazılımdır [22].

Şekil 2.7’de çeşitli ağ analiz yazılımları ve bunların birbirleriyle ilişkileri çizge yapısında gösterilmiştir [23].



Şekil 2.7. Ağ analiz yazılımları ağı (Pajek ile görselleştirilmiş)

2.3 Sosyal Ağ Analizindeki Aşamalar

Kullanılan veriler, çok geniş bir alandan toplanmaktadır. Bu veriler web sayfaları, log dosyaları, kullanıcı kayıt bilgileri, oturum ve hareket bilgileri ve site yapısı ve içeriğinden oluşmaktadır. Söz konusu büyüklükte ve karmaşıklıkta verilerin etkin bir biçimde analiz edilebilmesi için kullanılan tüm yöntemler 4 ana işlem basamağından oluşmaktadır: Kaynak bulma, bilgi çıkarımı ve ön işleme, genelleştirme ve çözümleme [24].

Veri Seçimi

Veri seçimi çok geniş kapsamlı bir konudur ancak kısaca bilginin elde edilme kısmıdır. Çeşitli verilerin çevrimiçi ya da çevrimdışı olmasına bakılmadan bir veri ambarında toplanmasıyla yapılmaktadır. İnternet üzerindeki çeşitli gazeteler, haber grupları vb. gibi yerlerden verinin toplanıp arama amaçlı olarak bir yerde saklanmasıdır.

Bilgi Çıkarımı ve Ön İşleme

Veri kaynağından toplanan verilerin işlenmesi ve işe yaracak hale getirilmesi, bir bakıma verinin temizlenmesidir. Ayrıca modellenmesi, sınıflandırılması hatta filtrelenmesidir.

Genelleştirme

Genelleştirmede, elde ettiğimiz tecrübenin genelleştirilmesi ve bu tecrübeler üzerinden genel-geçer kurallar türetilmesi amaçlanmaktadır.

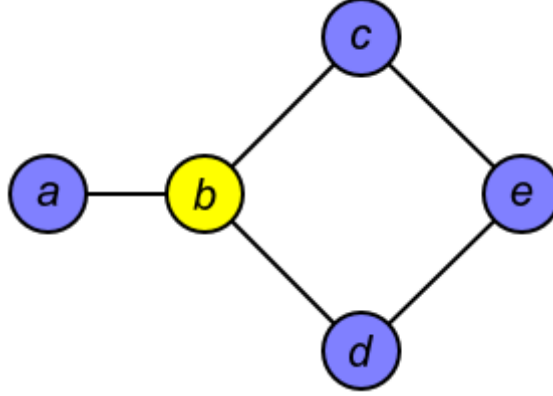
Çözümleme (Analiz)

Çıkarılan kurallar yardımı ile eldeki verilerden anlamlı sonuçlar çıkarılması ve ileriye yönelik tahmin ve değerlendirmelerin yapılmasıdır [18].

2.4. Sosyal Ağ Analizindeki Ölçütler

Arasındalık: Merkezilik ölçütleri içerisinde hesaplanması en karmaşık olan ölçüt olan arasındalık ölçütü, ağdaki bir düğümden ya da aktörden geçen en kısa yolların oranı ile bulunmaktadır. Şekil 2.8’de diğer düğümlerin arasında kalan B düğümü diğer düğümlerle ilişkiyi sağlamaktadır. Öncelikle ağdaki tüm düğüm çiftleri arasındaki en kısa yollar bulunmakta, ardından bu yolların kaçında o düğümün yer aldığı oranı arasındalık ölçütünü

vermektedir. Büyüklüğü fazla ağlarda hesaplanması çok maliyetli olabilecek bir ölçüt olduğu için belirli seviyedeki komşulara kadar inilerek de hesaplanabilmektedir. Arasındalık derecesi yüksek olan aktörler konumları dolayısıyla, diğer aktörlere göre daha önemli bir konumdadırlar ve ağda olup bitenden daha çok haberdar olmaktadır [25, 26].



Şekil 2.8. B düğümü A ile diğer düğümler arasında

Yakınlalık: Ağdaki herhangi bir düğümün, diğer tüm düğümlere olan en kısa ortalama uzaklıkların toplamının elde edilmesiyle bulunmaktadır. Bağlantıların yönlü olması durumunda kısa yolların bulunması aşamasında bu yönler dikkat edilmelidir. Bu durumda da giriş ve çıkış olmak üzere iki farklı yakınlalık ölçütü hesaplanmaktadır [25, 27].

Derece: Ağdaki bir aktörün derecesi, kendisine bağlı olan bağlantıların sayısı ile hesaplanmaktadır. Basit hesaplanmasının yanı sıra, aktörün önemini gösterebilecek önemli bir ölçüttür. Çoğu sosyal ağda, bir bireyin ne kadar çok bağlantısı var ise o kadar önemli ve güçlüdür gibi bir bakış açısı hakimdir. Hatta derecesi en yüksek olan aktör, ağın en aktif üyesi olarak da yorumlanabilmektedir. Bağlantıların yönlü olduğu durumlarda, giren bağlantıların sayısı ile çıkan bağlantıların sayısı ayrı ayrı hesaplanmaktadır [25, 26].

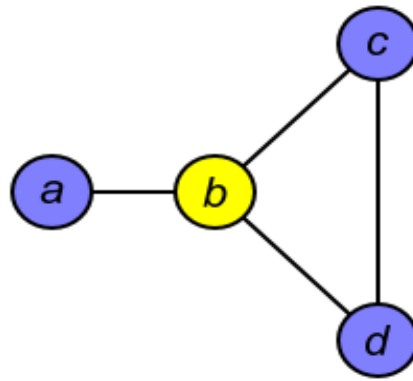
Merkezilik: Yakınlalık, arasındalık ve derecelendirme merkeziliğin ölçümüdürler. Bu düğümün sosyal gücünü ve ağla ne kadar iyi bir bağlantı kurduğunu göstermektedir. Merkezilik ile “Ağdaki en önemli ve merkez konumda bulunan aktör(ler) kim(ler)dir?” sorusuna cevap aranmaktadır. Ağdaki diğer düğümlerle en fazla hangi düğümün ilişkisi varsa o düğüm en merkezi düğümdür.

Köprü: Bağlantının kaldırılması diğer uç noktalara (düğümlere) ulaşılmamasına sebep olursa bu bağlantıya köprü denilebilir. Ağ grupları arasındaki tek bağlantıdır. Kaldırılması bu gruplar arasındaki bağlantıyı keser.

Merkezileştirme: Merkezileştirilmiş bir ağda daha derli toplu, dağınık olmayan analiz sonuçları alınabilmektedir. Çok merkezileşmiş bir ağda baskınlık birkaç merkezi düğümdedir. Bu merkezi düğümlere bir şey olursa ağ yapısı bozulabilir. Merkezi düğümler arasındalığı ve derecesi yüksek düğümlerdir. Ağ yapısında çok etkili düğümlerdir.

Yol Uzunluğu: Bir düğümün diğer düğümlerle arasındaki mesafeyi ölçmektir. Eğer bu iki düğüm arasındaki mesafe kısaysa bu düğümler arasındaki düğüm/kenar sayısı daha az demektir. Birbirine yakın düğümlerin ilişki kurma ihtimali daha fazladır.

Kümelenme Katsayısı: Kümelenme katsayısı bir aktörün, komşuları ile arasındaki bağlantının ne derece güçlü olduğu hakkında bilgi vermektedir. Bir aktörün komşularının kendi aralarındaki, gerçek bağlantıların olası tüm bağlantılara oranıyla hesaplanmaktadır [28]. “Arkadaşımın arkadaşı çoğunlukla arkadaşımdır” şeklinde düşünülürse, Eğer A düğümü B düğümlüne bağlıysa ve B düğümü C düğümlüne bağlı ise A düğümünün de C düğümlüne bağlı olma ihtimali yüksektir. Aşağıdaki Şekil 2.9’da BCD üçgeni oluşmuştur. B ’den geçme ihtimali olan üçgen sayısı 3’tür. Buna göre B düğümünün kümelenme katsayısı $1/3$ ’tür.



Şekil 2.9. B düğümü kümelenme katsayısı

Bütünlük: Hangi düğümlerin birbirleriyle güçlü bağlar oluşturduğunu ölçmektedir. Güçlü bağlantılar oluşturmuş düğümlerin benzerliği daha fazladır.

Yoğunluk: Ağda var olan tüm bağlantıların, olası tüm bağlantılara oranı yoğunluk ölçütünü vermektedir. Tüm aktörlerin birbirine bağlı olduğu ağlarda yoğunluk değeri 1 olacaktır. Yoğunluk değerinin yüksek olması ağdaki aktörlerin birbirlerine daha güçlü bağlı olduklarını ve birbirlerinden soyutlanmış durumda olmadıklarını göstermektedir. Yönlü bir çizgede yoğunluk $y = m/n(n - 1)$ olarak tanımlanmakta çünkü n bileşenin her birinden kalan $(n - 1)$ bileşenin her birine bir ilişki olabilmektedir. Yönsüz bir ilişkide ise $y = m/n(n - 1)/2$ olarak tanımlanmakta, çünkü iki bileşen arasında bir yöndeki ilişki ile diğer yöndeki aynı şeydir ve iki defa saymamak için formüldeki payda yarıya bölünmektedir. Tüm ağa dair yoğunluk ölçümü çok bilgilendirici değildir. Eğer aynı sosyal grubun yoğunluğu farklı zamanlarda ölçülüyorsa yoğunluktaki değişme bize bilgiler verebilmektedir. Fakat farklı iki sosyal grubu karşılaştırırken yoğunluk ölçüsü işe yaramaz çünkü büyük ölçüde grubun bileşen sayısına bağlıdır. Yani farklı büyüklükte gruplar arasında karşılaştırma yapmaya imkân vermez. Buna karşılık ağdaki tüm bireysel ağ yoğunluklarını incelemek faydalı bilgiler verebilmektedir. Bireysel ağ yoğunluğu ağın merkezi olan bireyi (ve ilişkilerini) dışarıda bırakarak onun ilişkili olduğu bireylerin arasındaki ağın yoğunluğu olarak hesaplanmaktadır.

Akış Arasındalık Merkeziliği: Tüm düğüm çiftleri arasında maksimum akış toplamına etki eden düğüm derecesidir.

Eigenvector Merkezilik: Derece hesaplanırken dikkate alınan aktörün sahip olduğu bağlantıların eşit olmadığını gösteren bir ölçüttür. Örneğin, ağdaki bir aktör için önemli aktörlere olan bağlantıların etkisi, diğer sıradan bağlantılardan daha fazla olabilmektedir. Bağlantıda olduğu aktörlerin daha merkezde olması, o aktörün daha merkezi bir konumda olacağını göstermektedir. Bu ölçüt hesaplanırken komşuların merkezilik derecelerinin toplamı dikkate alınmaktadır [27].

Lokal Köprü: Eğer bitiş noktası diğer komşularla paylaşılmıyorsa bu kenara lokal köprü denilebilmektedir.

Prestij: Yönlü çizgede düğümün merkeziliğini belirtmek için prestij terimi kullanılmaktadır.

Radiality: Bireyin ağından yeni bilgiler elde etme ve etkinlik ölçütüdür.

Ulaşım: Ağdaki herhangi bir üyenin derecesi ile diğer üyelere ulaşılabilirliktir.

Yapısal Bütünlük: Gruptan çıkarıldıklarında gruba bağlantının kesildiği minimum üye sayısıdır.

Yapısal Eşitlik: Bazı düğümlerin diğer düğümlere ulaşmasını sağlayan ortak bağlantı kümeleridir.

Her bir ölçüt ağ yapısı ile ilgili bilgiler vermektedir. Ağın analizinde ölçütlerin değerlerine göre çözümleme her seviyede yapılabilmektedir. Bütün bir sosyal sistemin ilişki örüntüleri incelenebileceği gibi, grup seviyesinde ya da bireysel seviyede de çözümleme yapmak mümkün olabilmektedir [16, 29, 30].

2.5. Sosyal Ağ Yapılı Verilerin Sınıflandırılması

Bağlantı madenciliğinde kullanılan temel birkaç alan bulunmaktadır

Düğüm Tabanlı

- Bağlantı Tabanlı Popülerlik
- Bağlantı Tabanlı Sınıflandırma
- Düğüm Kümeleme
- Düğüm Tanımlama

Bağlantı Tabanlı

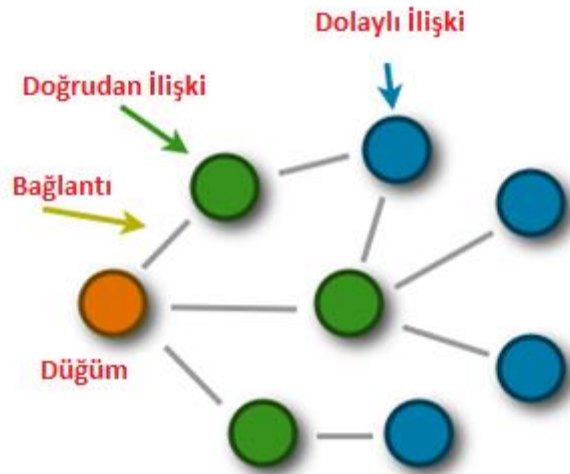
- Bağlantı Tahmini

Çizge Tabanlı

- Alt Çizge Keşfi
- Çizge Sınıflandırma
- Üretken Modeller [31].

3. BAĞLANTI TAHMİNİ

Sosyal ağlar bir yol haritası gibidir. Bireylerin birbirleriyle ilişki bağları kurmaları sayesinde oluşan bir yapı vardır. Bu yapı sayesinde oluşan veriler çizge kuramına göre yapılandırılabilir. Kişiler yani sosyal ağdaki üyeler birer düğüm olarak düşünülebilir. Şekil 3.1’de gösterildiği gibi aralarındaki bağlantı ve ilişki ise kenar olarak tanımlanabilir. Çizge kuramında olduğu gibi $G = (V, E)$ bağlantıları bir matris üzerinde gösterilebilir. Hangi üyenin kiminle ilişkisi olduğu bu bağlantı üzerinden tanımlanabilir.



Şekil 3.1. Sosyal ağın çizge yapısı

Sosyal ağlar dinamik, seyrek ve kolektif bir yapıya sahiptir. Bu yüzden üzerinde tahminlerde bulunmak zordur. Dahası erken evredeki popülerliğin durumuna göre, daha sonraki evrelerde de popülerliğin artarak devam edeceği olasılığı daha yüksek olabilmektedir. Fakat bu durumun tersinin olabileceği de tahminler arasında olması gerekmektedir. Bir etki alanındaki bağlantı veya ilişkilerin varlığının doğru tahmini önemli ve zor bir iştir. Diğer anlamda bu sosyal ağ üzerinde daha çok nelerin tahmini yapılabilir? Bu da önemli bir sorudur. Yaşadığımız bilgi çağı toplumunda bu ağlardaki ilişkiler sayesinde bireylerin sosyal davranışlarını inceleyebilmekte, insan ilişkileri konusunda nitel

ve nicel deęerlendirmelerde bulunabilmekte ve bunlardan ok yararlı bilgiler elde edilebilmektedir [30].

3.1. Baęlantı Tahmini Problemi

Sosyal aęlarda baęlantı tahmini problemi aynı zamanda aę yapısının tahmin edilmesi problemidir. Baęlantı tahmini yöntemlerinin erevesini dęüm bilgisi ve aę yapısı bilgisi oluřturmaktadır. Sosyal aęlardaki veriler ve iliřkiler izge yapısında dřünlmektedir. Veriler dęüm, iliřkiler ise baęlantı olarak tanımlanmaktadır. Dęümler bireyler ve baęlantı ise bu bireyler arasındaki iliřkilerdir.

Aę modelinde her dęüm vektrel yapılandırılmış veri olabilmektedir. Veri tabloları dnřtrlmekte, satırlar deęerleri stnlar ise zellik/nitelikleri belirtmektedir. Eęer tm dęmlerin kimlięi bilinir ve bazı baęlantı yapıları bilinirse buradan daha oluřmamıř bazı baęlantılar tahmin edilebilmektedir. Veya yeni dęmler bazı baęlantılar ile ilgili bilgi sunuyorsa ve bazı nitelikler belliyse buradan da yeni dęmler arasındaki baęlantılar tahmin edilebilmektedir. Aynı zamanda var olan baęlantının gelecekte de olup olmayacaęı da tahmin edilebilmektedir.

Bu tahminleri bilmek zor problemlerden biridir. Aę dinamik bir yapıya sahiptir. Var olan aę yapısının resmi ekilmektedir. Yakın gelecekte bu aęa yeni bireylerin dâhil olup olmayacaęının, yeni baęlantıların oluřup oluřmayacaęı veya var olan baęlantıların ileriki zamanlarda da devam edip etmeyeceęinin hesaplanması gerekmektedir. Aędaki bilgilerin nasıl tanımlanacaęı da nemli bir sorundur. Var olan bilgilerin nasıl kullanılacaęının, doęru hesaplamalar (tahminler) yapmada ne kadar etkili olacaęı iyi dřnlmelidir.

Baęlantı tahminin zorlukları  kısma ayrılabilir:

Birincisi, aęın topolojik bilgisine ek olarak, dęmler iin ekstra yan bilgi ve eř deęiřkenlere ihtiya duyulmaktadır. rneęin bir aęda protein iftleri arasındaki etkileřimde, her proteinin biyolojik zelliklerini tarif eden tanımlamalara ihtiya vardır. Bu bilgiler zellikle seyrek baęlı dęmlerin baęlantılarının tahmininde yararlı olabilmektedir. Bu řekilde farklı tipte bilgiler kullanılabilir. Tahminlerde daha performanslı sonular alınabilmektedir.

İkincisi, veri setleri ařırı dengesiz durumlar ortaya koyabilmektedir. Mevcut bilinen kenar (baęlantı) sayısı bilinenden daha az ıkabilmektedir. Bu durum dengeli verilerde kullanılan birok modelin etkinlięini engellemektedir.

Üçüncüsü, eğer geniş sayıdaki düğümler ve/veya kenarlar çizgede ölçeklenirse, hesaplama yapan modellerin verimli olması şarttır. Yani oluşturulan modelin büyük sistemlerde kullanılmasının zorluğu ön plana çıkmaktadır [32].

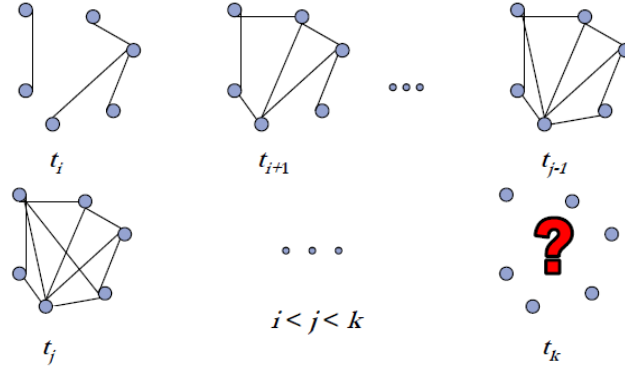
Bağlantı tahmini 4 farklı problemi işaret etmektedir. Bunlar;

- Bağlantının oluşumu
- Bağlantının türü
- Bağlantının ağırlığı
- Bağlantının önemi

Bu konuda yapılan çalışmaların büyük kısmı bağlantı oluşumu ile ilgilidir. Yani gelecekte sosyal ağda iki düğüm arasında bir ilişki kurulup kurulmayacağının tahmini yapılmaya çalışılmaktadır. Bu konuda daha çok çalışma yapılmasının sebebi ise bağlantının ağırlığı ve bağlantının önemi problemlerine kadar uzaması ve bunları da bir anlamda içinde barındırmasıdır. Bağlantının ağırlığı bağlantının ölçümünü, etkisini hesaplamaya çalışmaktadır. Bağlantının önemi benzer düğüm çiftleri arasında birden fazla bağlantı olup olmadığına odaklanmaktadır. Bağlantı türü ise iki nesne arasındaki ilişkinin türünü, rolünü anlamaya dayanmaktadır [31].

3.2. Matematiksel Tanımı

Bağlantı tahmini problemi şu şekilde belirtilebilmektedir. Düğümler V yani veri örnekleri olarak düşünülürse $V = \{v_i\}_{i=1}^n$, E ise bu veriler üzerinde var olan ilişkileri göstermektedir. Buna göre sosyal ağ $G = (V, E)$ olarak tanımlanabilmektedir. Buradan (v_i, v_j) düğüm çiftleri olacaktır. Bu düğüm çiftleri arasında, $e_{ij} \notin E$ oluşmamış bağlantının tahmini yapılmaya çalışılmaktadır. Düğüm çiftleri arasındaki bağlantının önemini gösteren ölçüt de $skor(x, y)$ fonksiyonu olarak tanımlanabilmektedir. Daha sonra bu fonksiyon değerleri büyükten küçüğe sıralanmaktadır. Bu şekildeki hesaplamalar ve ölçümlerle iki düğüm arasındaki benzerlik veya yakınlık bulunabilmektedir. İki düğüm arasındaki en kısa yol veya bağlantı durumlarına göre ilerde bağlantı oluşma ihtimali gibi hesaplamalar yapılabilmektedir [30]. Şekil 3.2’de oluşan ağın zamansal değişimi yapılan ölçümlerle karşılaştırılarak tahminin doğruluğu ölçülebilmektedir.



Şekil 3.2. Sosyal ağın zamansal değişimi [16].

3.3. Yapılan Bazı Çalışmalar

Bu konuda yapılan bazı çalışmaları ele almak yararlı olacaktır.

- İlişkisel yapıları keşfetme, kümeleme
- Bağlantıları kullanarak varlıkların sınıflarını/özelliklerini tahmin etme
- Bilinen varlık sınıflarında bağlantı türünü tahmin etme
- Yüksek boyutlu uzayda konumlanmış bağlantıları tahmin etme
- Tekli çizge tabanlı özellik kullanarak potansiyel bağlantıları sıralama [30].

Daha da detaya inip birkaç örnek verilebilir [33];

- Seglen [34] geniş etkisi olan faktörlerle dergilerdeki makalelerin eğilimlerini analiz etti.
- Vinkler ve Davidson [35] büyüyen dergilerdeki makalelerin atıf yapılma olasılığının daha yüksek olduğunu gösterdiler.
- Liben-Nowell ve Kleinberg [36] büyük ortak yazar ağları için bağlantı tahmini modeli önerdiler.
- Popescul ve Ungar [37] İstatistiksel İlişkisel Öğrenme için bağlantı tahmini modelini inşa eden yeni bir yaklaşım önerdiler.
- Hasan ve arkadaşları [3, 38] birkaç denetimli öğrenme yöntemini bağlantı tahmini için test ettiler. Bu modeller karar ağaçları, k-en yakın komşu, çok katmanlı

algılama, destek vektör makinesi, radyal tabanlı fonksiyon ağı ve bunlar arasındaki en başarılı sonuçlar Destek Vektör Makinesi modeli ile olmuştur.

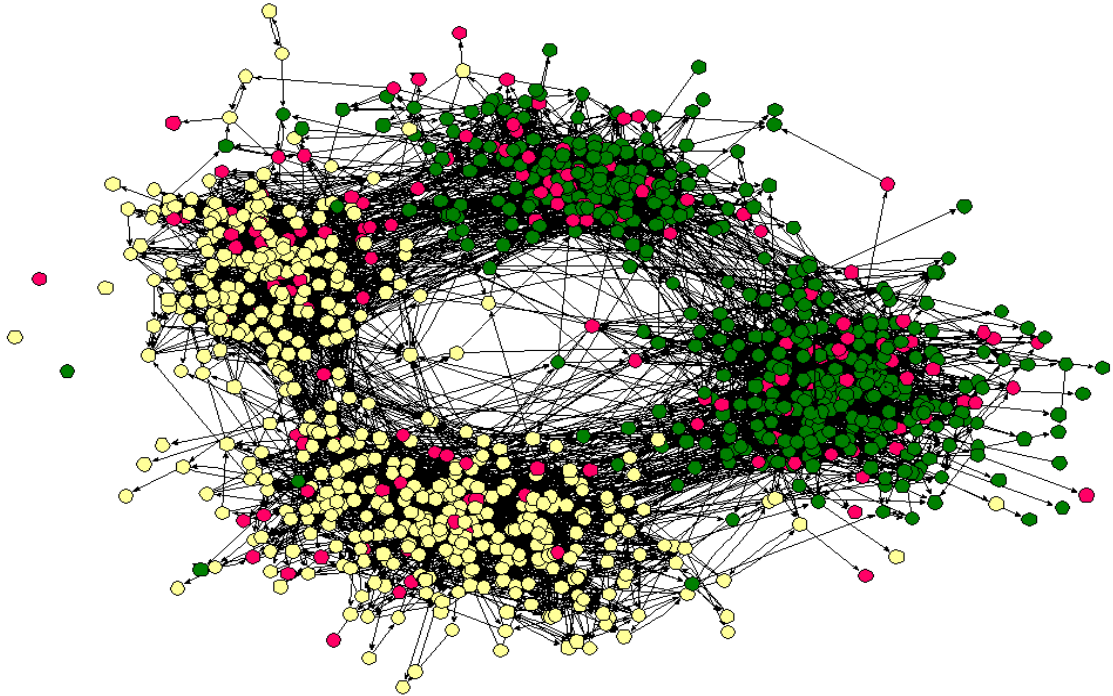
- Clauset ve arkadaşları [39] sosyal ağların hiyerarşik yapısını araştırarak kısmen bilinen ağların eksik bağlantılarını yüksek doğruluk ile tahmin ettiler.
- Murata ve Moriyasu [40] İlan Tahtaları Soru Yanıtlama sosyal ağlarına Liben-Nowell ve Kleinberg modelini uyguladılar.
- Caragea ve arkadaşları [41] kümelenme yaklaşımına dayalı Live-Journal'da potansiyel arkadaşlıkları tahmin eden bir algoritma geliştirdiler. Live-Journal kullanıcı etkileşimlerine dayalı bir sosyal ağ günlüğü servsidir.
- Lu ve arkadaşları [42] iki düğüm arasında bir bağlantı varlığı olasılığını tahmin etmek için yerel bir yol endeksi sundular.
- Hwang ve arkadaşları [1] ortak yazarlı ağlar için tavsiye motorları önerdiler.
- Davis ve arkadaşları [43, 44] hastanın geçmiş bilgilerinden hastalık tahmini yapmak için yöntemler önerdiler.
- Folino ve arkadaşları [4] hastalık riskini tahmin eden bir yaklaşım sundular.
- Shibata ve arkadaşları [33] farklı sosyal ağlar için farklı öğrenme yaklaşımları ve farklı özelliklerle çeşitli bağlantı tahmini modelleriyle atıf tahmini önerileri yaptılar.
- Folino ve Pizzuti [5] hastalık ağlarında bağlantı tahmini yöntemlerini kullanan yaklaşımlar sundular.

3.4. İlişki ve Ağ Türleri

Sosyal ağlarda sevmeye, evlilik, profesyonel tavsiye verme, vb. birçok ilişki anlamlandırılabilir. Mesela Şekil 3.3'te ki etnik kökenlerine göre bir okuldaki arkadaşlık ağı bir çok anlam barındırmaktadır [45]. Bu çeşitliliğe bakıldığında bunlar belirli ilişki kategorilerine ve teknik özellik profillerine göre sınıflandırılabilir. Burada kısmen Borgatti ve arkadaşlarının [46] çalışmalarından elde edilen sınıflandırmaları kullanmak uygulamaya göre seçilecek yöntemler konusunda yardımcı olacaktır. Bu sınıflandırmanın özeti aşağıdaki Tablo 3.1'de verildi [47].

Tablo 3.1. Sosyal ağ arařtırmalarında karřılařılan baē t rleri ve veri seti  zellikleri

Baē t�r�	�rnekler	�izge/veri seti �zellikleri
Benzerlikler	Aynı yerde yařama Aynı kul�p/etkinliēe gitme Aynı cinsiyetten olma Benzer alışkanlıkları olma	İki mod ve y�nl� kullanmadan �nce d�n�řt�rme gerekir d�n�řt�rme sonucu y�ns�z iliřki �oēu zaman aēırlıksız
İliřkiler	Akrabalık, evlilik, arkadaşlık, iř iliřkisi, Sevme/sevmeme	�oēu zaman y�ns�z �oēu zaman aēırlıksız
Etkileřimler	G�r�řme, yardım etme, tavsiye verme	Genellikle y�nl� ve aēırlıksız
Akışlar	Bilgi akışı, řirketler arası personel akışı, �lkeler arası ticaret	Genellikle y�nl� ve aēırlıklı



řekil 3.3. Amerika'daki bir okuldaki arkadaşlık aēı. Renkler etnik k kenlerini g stermektedir.

3.5. Benzerlik İlişkileri ve İki-Mod Ağlar

Benzerlik ilişkileri esasen gerçek ilişkiler değildir. Yine de aynı mahallede yaşama, benzer kötü alışkanlıklar, vb. benzerlikler sosyal araştırmalarda kullanılabilmektedir. Bu bağlar çoğu zaman gerçek bağlardan (örneğin arkadaşlık) oluşan ağlarla ne kadar örtüşüklerini karşılaştırmak için kullanılabilmektedir. Öte yandan aynı "kulübe üye olma" gibi bağlar gerçek ilişkilere daha yakındırlar, çünkü aynı kulübe üye olan insanların birbiriyle büyük olasılıkla az ya da çok bir etkileşimi olacağı varsayılabilmektedir.

Bu kategoriyle ilgili bir örnek Galaskiewicz [48] tarafından derlenmiştir. Bu veri seti büyük şirket yöneticisi 26 kişi ve eşlerinden oluşan bir örneklemin hem şirket yönetim kurullarına hem de şehir kulüplerine üyelik verilerini içermektedir. Bu örnekte söz konusu kurul ve kulüplerin niteliği sebebiyle "aynı kulübe üye olma" durumunun gerçekten de iki kişi arasında bir etkileşime karşılık gelme olasılığı çok yüksektir.

Bu veri setinin bir teknik özelliği var: iki-mod. Veri setinde bir yanda bireyler bir yanda da kulüpler var ve ilişkiler her zaman birinci setteki bileşenlerden ikinci setteki bileşenlere doğru olmaktadır. Dolayısıyla sosyal ağ veri setini oluşturan bileşenler ilişki açısından tamamen farklı iki grubu oluşturmaktadır. İki-mod ismi bu yüzden kullanılmaktadır. Çizgelerin üç-mod, vb. olması da mümkündür, ancak sosyal ağlarda böyle bir durumla pek karşılaşmamaktadır.

Sosyal ağ analizinde yapılan analizlerin çoğunluğu tek-mod ağlara uygundur. Bu yüzden iki-mod veri setlerinin çoğu zaman tek-mod'a dönüştürülmesi gerekmektedir. Galaskiewicz [45] veri setinin toplanmasındaki amaç bireyler arasındaki ilişkiyi kulüpler üzerinden dolaylı ortaya çıkartmaktır. Böyle bir dönüştürme matris cebiri ile yapılabilir. Bir örnek üzerinde incelenirse iki-mod matris aşağıdaki şekildeki gibi olabilir.

	Kulüp1	Kulüp2
Birey1	1	1
Birey2	1	0
Birey3	0	1

Şekil 3.4. İki mod ilişki matrisi

Bu örnekte Birey1'in Birey2 ile Kulüp1 üzerinden ve Birey3 ile Kulüp2 üzerinden ilişkisi olduğu görülebilmektedir. Bu matris 3x2 boyutlarında, çünkü iki-mod sosyal ağda bir bileşen seti 3 bireyden, diğer bileşen seti ise 2 kulüpten oluşmaktadır. Bunu matris cebiri ile Şekil 3.5'deki gibi birey-birey ilişkisi matrisine dönüştürmek için oluşan matris devrik matrisi ile çarpılmaktadır.

$$\begin{array}{|c|c|} \hline 1 & 1 \\ \hline 1 & 0 \\ \hline 0 & 1 \\ \hline \end{array} \times \begin{array}{|c|c|c|} \hline 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline 2 & 1 & 1 \\ \hline 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \hline \end{array}$$

Şekil 3.5. İki moddan tek moda dönüştürme

Ortaya çıkan 3x3 matris artık bireyler arası ilişki matrisidir. İşlem ters yapılırsa, yani devrik matris ile orijinal matris çarpılırsa bu sefer 2x2 boyutunda kulüpler-kulüpler ilişkisini veren bir matris oluşacaktır.

Bulunan birey-birey matrisinin tek problemi var bireyler kendileriyle ilişki içerisinde görünmektedirler. Bunu gidermek için diyagonalı temizlemek gerekmektedir. Sonuç olarak Şekil 3.6'daki matris bulunmaktadır [47].

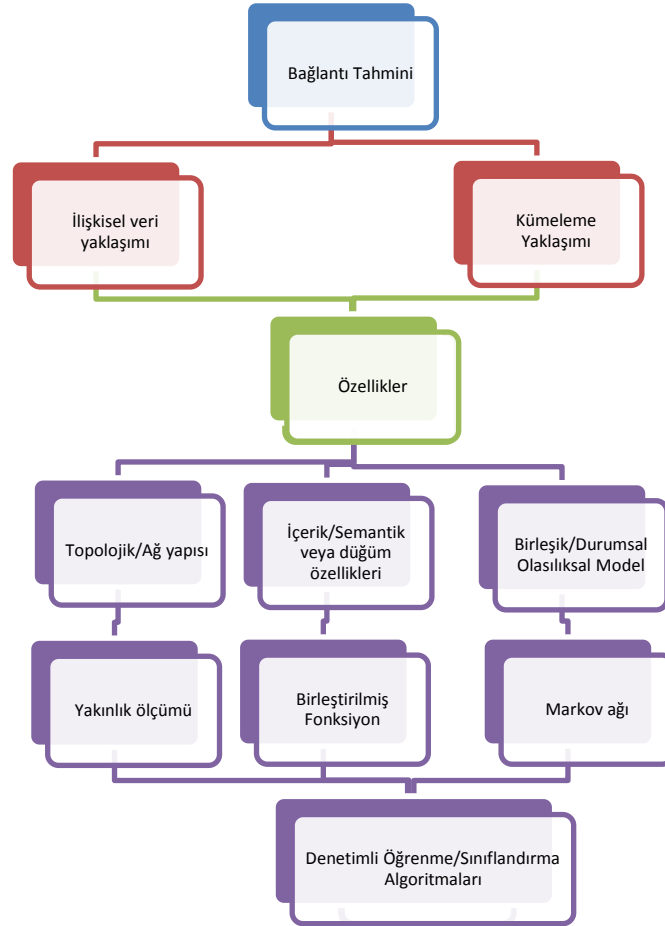
	Birey1	Birey2	Birey3
Birey1	0	1	1
Birey2	1	0	0
Birey3	1	0	0

Şekil 3.6. Komşuluk matrisi

Yapılan uygulamada da 2-mod olan hastalık-ilaç ağı da yukarıda açıklanan yöntemler kullanılarak tek-moda dönüştürüldü.

3.6.Bağlantı Tahmini Yöntemleri

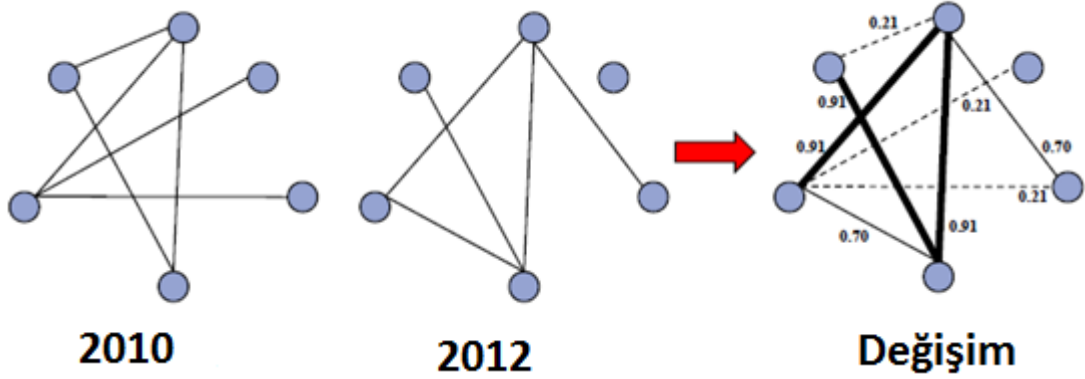
Bağlantı analizi ve sosyal ağların analizi yöntemleri verileri tek bir ilişkisel tablo olarak varsayar (Ör. Market-Sepet Analizi). Bu tür veri gösterimine örnek, müşteriye ürün satışlarında yeni önermeler yapmak gösterilebilmektedir. Diğer taraftan bağlantı tahmini ile birbirine bağlı birden çok ilişkisel tablodan yeni bilgilerde önerilebilmektedir.



Şekil 3.7. Bağlantı tahminine farklı yaklaşımlar [49].

Şekil 3.7’de gösterildiği gibi bağlantı tahmini yapılırken ağ yapısı veya problemin tanımlanmasına göre çözüme gitmek için farklı yaklaşımlar geliştirilebilir. Çeşitli verilerden bağlantı tahmini yapılmaya çalışılırsa; Ağ da iki varlık olduğu varsayılır ve iki varlık arasında ilişki kurulup kurulmadığı tanımlanmaya çalışılırsa, sınıflandırma varlıkların özellikleri ve çizge tabanlı dolaylı ilişki özellikleri üzerinden yapılabilmektedir. Bu özellikler, komşu sayısı, ilgi alanları, başlık, iştirakleri, demografik veriler (coğrafi konum gibi) olabilmektedir. Çizge yapısında düşünüldüğünde en kısa yol uzunluğu,

komşuluk yoğunluğu, göreceli önemi, ortalama ilk geçiş süresi bilgileri önem kazanmaktadır. Bayes ağları gibi yönlü grafiksel model veya yönsüz grafiksel model (ör. Markov Ağları) özelliklere göre bağımlı ilişkileri rahatça yakalar ve olasılıksal tahminler yapılmasını mümkün kılar. Şekil 3.8’de görüldüğü gibi farklı tarihteki ilişkilere göre yeni ilişkilerin oluşma ihtimali hesaplanabilmektedir.



Şekil 3.8. Ağ yapısındaki ilişkilerin zamansal değişimi [30].

Sosyal ağlarda ilişki tahmini için kullanılan temel iki yöntem vardır. Bunlar makine öğrenmesinde kullanılan Denetimli ve Denetimsiz yöntemlerdir.

Denetimli yöntemler yapısal özellikleri ayıklayarak eşleme fonksiyonu kullanarak sonuç elde etmeye çalışmaktadır. Eldeki çeşitli verilerden kıyaslama yaparak kararlar vermektedir. Belirli bir düğüm çifti arasında bağlantı olup olmadığını tahmin edebilecek bir ikili sınıflandırıcı öğrenmeye dayanmaktadır.

Denetimsiz yöntemler ise çeşitli benzerlik bulma yöntemlerini kullanmaktadır. Ayrıca, bir bağlantının var olup olmadığını sınıflandırmak için karar ağaçları, k -en yakın komşu veya destek vektör makineleri gibi çeşitli denetimli öğrenme/sınıflandırma algoritmaları kullanılabilir.

Bağlantı tahmininde denetimsiz öğrenme yöntemleri ile büyük veriler arasında daha az sistem gereksinimi duyarak bağlantı tahmini yapmak daha mantıklıdır. Bu yüzden benzerlik tabanlı algoritmalar daha iyi sonuçlar vermektedir. Aşağıda benzerlik tabanlı algoritmalar açıklanmaktadır [30].

x ve y düğüm çiftleri arasındaki bağlantıya göre $skor_{xy}$ atanır. Bu değer x ve y arasındaki benzerliği ölçer. Bu skorlara göre bağlantılar sıralanmakta ve en yüksek skorlara

sahip düğümler arasında benzerlik fazla olacağı için bağlantı oluşma ihtimali artmaktadır. Basitliğine rağmen, üzerinde yapılan çalışmalar açısından sorunlu bir alandır. Düğümlerin benzerliğinin tanımlanması sıkıntılı olabilmektedir. Benzerlik indeksi çok basit olabileceği gibi çok karmaşık da olabilmektedir. Bazı ağlarda iyi sonuç verebileceği gibi, bazılarında ise hiç sonuç alınamayabilir. Düğümlerin gerekli bazı özelliklerini kullanarak düğüm benzerliği tanımlanabilir. Bu düğümlerin benzer özelliklerinden düğümlerin benzerlikleri ölçülebilir. Bazen düğümlerin nitelikleri gizlidir. Böyle durumlarda yapısal benzerlik dediğimiz ağ yapısı ile ilgili benzerlik indekslerine odaklanılır. Yapısal benzerlik indeksleri çeşitli yollarla sınıflandırılır. Bunlar yerel-genel, parametre bağımlı-parametre bağımsız, düğüm bağımlı-yol bağımlı sınıflandırmalardır. Benzerlik indeksleri yapısal eşitlik ve düzenli eşitlik olarak sınıflandırılabilir. Bu konuda yapılan iki benzerlik varsayımı vardır. Birincisi bağlantının kendisi iki uç arasında benzerlik olduğunu gösterir, ikincisi komşuları benzer olanlar arasında benzerlik olacağı varsayımdır. Aşağıda indeksleme yöntemlerinden bazıları açıklandı [50].

3.6.1 Yerel Benzerlik İndeksleri

Ortak Komşu : $\Gamma(x)$ x 'in komşularının ve $\Gamma(y)$ y 'nin komşularının kümesi olsun. Bu durumda x ve y düğümlerinin birçok ortak komşusu varsa bu iki düğüm arasında bağlantı vardır veya oluşma ihtimali vardır denilebilir.

$$S_{xy} = |\Gamma(x) \cap \Gamma(y)| \quad (3.1)$$

Bu özellik kullanılarak, ortak komşu sayısını hesaplama, muhtemel benzerlikleri ortaya çıkarma gibi bir çok farklı çalışmada kullanılabilir. Açıklanan yöntemlerden bazıları da farklı normalizasyon metotları ile ortak komşuluğu kullanmaktadır.

Salton İndeksleme: k_x , x düğümünün derecesi, k_y ise y düğümünün derecesini göstermektedir. Buna literatürde kosinüs benzerliği de denmektedir.

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x + k_y}} \quad (3.2)$$

Jaccard İndeksleme: Ortak komşular toplam komşulara bölünerek sonuç elde edilmektedir.

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (3.3)$$

Sorensen İndeksleme: Bu indekislemede ekolojik topluluk verileri için kullanılmaktadır.

$$S_{xy} = \frac{2|\Gamma(x) \cap \Gamma(y)|}{k_x + k_y} \quad (3.4)$$

Hub Promoted İndeksleme: Bu indeks metabolik ağlarda yüzey çiftlerinin örtüşme miktarının hesaplanması için önerilen bir yoldur.

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min\{k_x, k_y\}} \quad (3.5)$$

Payda düşük tutulduğu için merkeze bitişik bağlantılar muhtemelen yüksek puan alacaktır.

Hub Depressed İndeksleme: Bunda ise yukarıdaki indekislemenin tersi bir sonuç alınacaktır.

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max\{k_x, k_y\}} \quad (3.6)$$

Leicht-Holme-Newman İndeksleme: Mümkün olmayan en yüksek değerle karşılaştırıldığında, birçok ortak komşusu olan düğüm çiftlerine yüksek benzerlik atanır, ancak beklenen komşu sayısından daha fazla olur.

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{k_x \times k_y} \quad (3.7)$$

Tercihli Bağlantı İndeksleme: Ölçeksiz büyüyen ağlarda kullanılmaktadır. x düğümüne yeni bir bağlantının oluşturulma olasılığı k ile orantılıdır. Aynı yapı büyüme olmayan bir ağda belirli zaman dilimlerinde eski bir bağlantının kaldırılması ve yeni bir bağlantının

oluşturulması olasılığı k ile orantılı yapılır. x ve y düğümleri arasında bağlantı oluşma olasılığı $k_x \times k_y$ ile orantılı olur.

$$S_{xy} = k_x \times k_y \quad (3.8)$$

Adamic-Adar İndeksleme: Bu indeksleme az bağlantılı komşulara daha fazla değer vererek ortak komşuları hesaplar. k_z , ortak komşu z düğümünün kaç düğümle bağlantısının olduğunu yani derecesini verir.

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z} \quad (3.9)$$

Kaynak Tahsisli İndeksleme: Kompleks ağlarda kullanılır. Birbiriyle direk bağlantılı olmayan düğüm çiftleri arasındaki bağlantıları ölçer. Birbirleri arasında bağlantı olmadığı halde düğümler ortak komşuları arasında iletim sağlarlar. İletim sağlayan her düğümün kaynak ünitesi vardır ve bu kaynağı komşularına eşit dağıtır. Bu düğümler arasındaki benzerlik birbirlerinden aldıkları kaynaklara göre hesaplanır.

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \quad (3.10)$$

Liben-Nowell ve arkadaşları [36] ve Zhou ve arkadaşları [42] yerel benzerlik indislerini kullanarak gerçek ağlar üzerinde kıyaslamalar yaptılar. Sonuçları Tablo 3.2’de verilmiştir [50].

Tablo 3.2. Lokal benzerlik algoritmaları karşılaştırması

İndisler	PPI	NS	Grid	PB	INT	USAir
Ortak Komşu	0,889	0,933	0,590	0,925	0,559	0,937
Salton	0,869	0,911	0,585	0,874	0,552	0,898
Jaccard	0,888	0,933	0,590	0,882	0,559	0,901
Sorensen	0,888	0,933	0,590	0,881	0,559	0,902
Hub Promoted	0,868	0,911	0,585	0,852	0,552	0,857
Hub Depressed	0,888	0,933	0,590	0,877	0,559	0,895
Leicht-Holme-Newman	0,866	0,911	0,585	0,772	0,552	0,758
Tercihli Bağlantı	0,828	0,623	0,446	0,907	0,464	0,886
Adamic Adar	0,888	0,932	0,590	0,922	0,559	0,925
Kaynak Tahsisli	0,890	0,933	0,590	0,931	0,559	0,955

Tablo 3.2’de 6 farklı ağlar kullanılarak yukarıda tanımlarını verdiğimiz yerel benzerlik indisleri karşılaştırılmaktadır. Ağlarımız **PPI**: Protein-Protein etkileşim ağı, **NS**: Ortak yazar bilimsel yayın ağı, **Grid**: Batı Amerika elektrik şebekesi ağı, **PB**: Amerika Politika blogları ağı, **INT**: router seviyesi internet ağı **USAir**: ABD hava taşımacılığı ağı’dır. [36, 42]. Karşılaştırmalarda en kötü sonuçları *Tercihli Bağlantı* indekseleme vermiştir. Fakat *Tercihli Bağlantı*’nın en büyük avantajı en az bilgiyle sonuçlar üretebilmesidir. En iyi sonuçları da *Kaynak Tahsisli* indekseleme vermiştir. İkinci en iyi performansları da *Adamic Adar* ve *Ortak Komşu* indekseleme vermiştir. Yapılan hastalık-ilaç bağlantı tahmini uygulamasında *Ortak Komşu*, *Adamic Adar*, *Jaccard* ve *Kaynak Tahsisli* yöntemleri kullanıldı.

3.6.2 Genel Benzerlik İndeksleri

Katz İndeksleme: Yolları gruplandırma yöntemi üzerine kurulmaktadır. Yol uzunlukları toplanmaktadır. En kısa yollara daha fazla ağırlık verilmektedir.

$$S_{xy} = \sum_{l=1}^{\infty} \beta^l \cdot |\text{paths}_{xy}^{(l)}| = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \dots, \quad (3.11)$$

$\text{paths}_{xy}^{(l)}$ yolların kümesi, l ise x ve y ’yi birbirine bağlayan uzunluktur. Yolun ağırlığını kontrol etmek için kullanılan β serbest değerler alan bir parametredir. Benzerlik matrisi:

$$S^{Katz} = (I - \beta A)^{-1} - I \quad (3.12)$$

Leicht-Holme-Newman İndeksleme (LHN2): Katz indekslemenin bir varyasyonudur. Eğer iki düğümün komşuları benzerse düğümler birbirine benzerdir mantığıyla formüllendirilmektedir.

$$S = \emptyset AS + \varphi I = \varphi(I - \emptyset A)^{-1} = \varphi(I + \emptyset A + \emptyset^2 A^2 + \dots), \quad (3.13)$$

$\emptyset$ ve φ özel parametrelerdir. φ , 1 değerini alınca Katz ile benzer sonuçlar verir.

Ortalama İşlem Süresi: Ortalama bir düğümden diğer düğüme gidip gelme süresini hesaplayarak benzerlik ölçümü yapılır. $m(x, y)$, x düğümünden y düğümüne rastgele gidiş adımıdır.

$$n(x, y) = m(x, y) + m(y, x) \quad (3.14)$$

Kosinüs Benzerliği: İki düğüm arasındaki kosinüs uzaklığını bularak bu düğümlerin birbirlerine benzerliği ölçülür. Kosinüs benzerliği 0 ile 1 arasında bir değer alır. Kosinüs benzerliği değerinin 1'e yaklaşması birbirleriyle olan benzerliklerinin daha fazla olduğu anlamına gelir.

$$S_{xy} = \cos(x, y) = \frac{l_{xy}}{\sqrt{l_{xx} + l_{yy}}} \quad (3.15)$$

Yeniden Başlamalı Rastgele Yürüyüş (RWR): X düğümünden rasgele hareket edip belli bir c olasılığı ile komşuya gitmesi ve $1 - c$ olasılığı ile tekrar x düğümüne dönmesidir.

$$\vec{q}_x = cP^T \vec{q}_x + (1-c) \vec{e}_x \quad (3.16)$$

P geçiş matrisidir. Eğer x ve y düğümleri arasında bağlantı varsa $P_{xy} = 1/k_x$, yoksa $P_{xy}=0$. $S_{xy} = q_{xy} + q_{yx}$, q_{xy} vektör $\vec{q}_x$ in y 'inci elemanıdır .

Simrank: Benzer düğümlere bağlanıyorlarsa, düğüm çifti benzerdir. *Rasgele Yürüyüş*'te olduğu gibi x ve y düğümlerinden hareketle hangi düğümde karşılaşacakları hesaplanmaktadır.

$$S_{xy} = c \frac{\sum_{z \in \Gamma(x)} \sum_{z' \in \Gamma(y)} S_{zz'}}{k_x \cdot k_y}, S_{xx}=1 \text{ ve } c \in [0,1] \quad (3.17)$$

Matrix Forest İndeksleme (MFI): x ve y arasındaki benzerlik bir ormanda aynı ağaç kökünden gelmeleri iki düğümü benzer yapar mantığı vardır.

$$S = (I + L)^{-1} \quad (3.18)$$

3.6.3 Yarı Yerel İndeksler

Yerel Yol İndeksleme: Ortak komşu indekslemeye benzer. Farkı daha karmaşık ve zor hesaplamalarda ϵ parametresi kullanmasıdır.

$$S = A^2 + \epsilon A^3 + \epsilon^2 A^4 + \dots + \epsilon^{n-2} A^n, \quad (3.19)$$

Yerel Rastgele Yürüyüş (LRW): Yerelde rastgele bir düğümden başlanarak düğümlerin dolaşımından bağlantının hesaplanmasıdır.

Çakıştırılmış Rastgele Yürüyüş (SRW): RWR indeksleme ile benzerdir, fakat tekrar ilk düğüme dönüşte hedef düğümle en yakın düğüm arasında en yüksek benzerlikle sonuçlandırılmaktadır.

$$S_{xy} = \sum_{r=1}^t [q_x \pi_{xy}(r) + q_y \pi_{xy}(r)], \quad (3.20)$$

4. UYGULAMA SONUÇLARI

Sağlık Bakanlığımızın 6 ilde toplam 5 bin reçete üzerinde yaptığı ve yayınladığı çalışmada ortaya çıkan sonuçlar yaptığımız uygulamanın önemini göstermektedir. Yapılan çalışmada bademcik iltihabı teşhisi ile yazılan reçetelerin yüzde 86'sının rasyonel olmadığı tespit edilmiştir [51]. Dünya Sağlık Örgütü'ne göre dünya genelinde akılcı olmayan reçete oranı ise yüzde 50'ler civarındadır [52].

İnsan sağlığının önemini düşündüğümüzde bu sonuçlar gerçekten düşündürücüdür. Yukarıda açıklanan akılcı olmayan reçete oranları maddi ve manevi birçok zarara sebep olmaktadır. Günümüzde teknolojiadaki gelişmeler bu sıkıntıyı önemli oranda azaltabilir. Bu oranları aşağıya düşürmek için çeşitli çalışmalar yapılabilir. Bu çalışmalardan biride doğru ilaçların doğru hastalıklara yazılmasını sağlayacak öneri sistemlerinin, yönlendirmelerin yapılabileceği yazılımların bu sistemlerde kullanılmasıdır.

Yukarıdaki sebeplerden yola çıkılarak sosyal ağ analizi yöntemlerini kullanarak hastalıklara doğru ilaçların seçilmesi, ilaç ve/veya hastalık benzerliklerinin bulunması ve hastalıklara doğru reçetelerin yazılması için ilaç ile hastalık arasındaki ilişkiden bağlantı tahmini yapıldı. İçinde hastalık ve ilaç bilgilerinin bulunduğu yapay bir veritabanı oluşturuldu. Doktora tedavi olmaya gelen hastaya teşhis konularak, bu teşhise göre daha önceden belirlenen ilaçlar arasından hastalara reçete verilmesi senaryosu kullanıldı. 10 hastalık ve bu hastalıklarda tedavi için kullanılan 5'er adet ilaç belirlendi. Veritabanında kullanılan hastalıkların seçiminde belli bir sebep aranmadı. Rastgele hastalıklar seçildi. Farklı hastalıklar olması ve bir hastalığın altındaki farklı alanlar seçilmediği için ilaçların birbirine yazılması uygun görülmediğinden doğrudan hastalık için seçilen ilaçlar birbirine yazılmadı. Aynı zamanda karmaşık bir yapı olmasını engellemek için ilişki ağrı kesici ilaçlar üzerinden oluşturuldu. Bu yüzden 5 adet ağrı kesici ilaç belirlendi. Böylece ortak ilaçların yazılması sebebiyle hastalıklar ve ilaçlar arasında bağlantı oluştu. İlişkiler bu ağrı kesici ilaçlar üzerinden oluştu. Ağrı kesici ilaçlar hastalıklara yazılırken belli bir kriter yoktu. Yani rastgele veritabanında hastalıklara yazıldı. İlaçların bu hastalıklara uygun olup olmadığı internetten araştırılarak belirlendi. İlişkiler ile ilgilenildiği için ilaçların hastalıklara uygunluğu çok önemli değildir. Önemli olan ilişki üzerinden bağlantı tahminin yapılmasıdır.

Veritabanı, Tablo 4.1’deki 10 hastalık ve bu hastalıklara yazılan 55 adet ilaç olan bir yapay veritabanından oluşturuldu.

Tablo 4.1. Veritabanında kullanılan hastalıklar ve ilaçlar

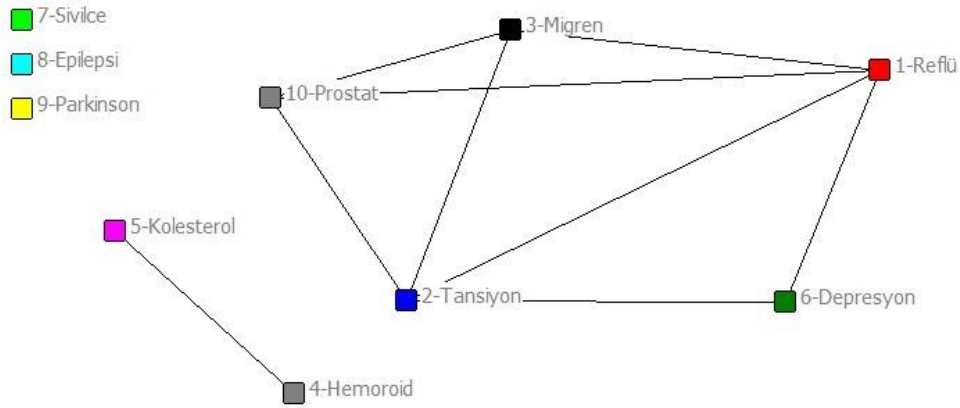
Hastalık	İlaç1	İlaç2	İlaç3	İlaç4	İlaç5
1-Reflü	Abpax	Acilibre	Antepsin	Aprazol	Arinna
2-Yüksek Tansiyon	Accuzide	Aceper	Aceprix	Acuitel	Adalat
3-Migren	Avmigran	Cafergot	Dopergin	Ergafein	Gravis
4-Hemoroid	Doloproct	Hedensa	Hemoralgine	Kortos	Proctolog
5-Kolesterol	Alvastin	Ateroz	Ator	Atopir	Avitorel
6-Depresyon	Anafranil	Antimax	Aplenzin	As-Cilog	As-Sertral
7-Sivilce	Acnegen	Acnelyse	Aknefug	Aknetrent	Aknilox
8-Epilepsi	Alond	Alyse	Apilep	Convulex	Daylep
9-Parkinson	Akineton	Artane	Azilect	Cabaser	Comtan
10-Prostat	Avodart	Dilaprost	Dutapros	Dynasef	Flomax
Ağrı Kesici İlaçlar	Panalgin	Minoset	Aspirin	Vermidon	Novalgin

Oluşturulan bu yapay veritabanını ağ yapısına dönüştürmek için birbirleriyle ilişkileri üzerinden komşuluk matrisi oluşturuldu. İki mod olan hastalık ilaç ağı bağlantılara göre komşuluk matrisinde hastalıklara ilaç yazılmışsa 1, yazılmamışsa 0 değeri aldı.

Şekil 4.1’de matris üzerinden oluşan ağ görülmektedir. Beraber kullanılan ortak ilaçlar üzerinden hastalıklar arasında bağlantı olduğu görülmektedir. Ortak kullanılan ilaçlar ağrı kesici ilaçlardır. Yuvarlak ve kırmızı renkteki şekiller hastalıkları, renkli karelerde ilaçları belirtmektedir. Düzgümler arasında ilişki, kahverengi ve kare şeklinde olan ağrı kesici ilaçlar üzerinden oluşmaktadır. En çok farklı hastalıklarda kullanılan ortak ağrı kesici *Vermidon* ilacıdır. En çokta *Panalgin* ile beraber kullanıldı. *Vermidon* ilacı Reflü, Yüksek Tansiyon, Migren, Prostat hastalıkları için yazıldı. Düzgüm derecesi 4’tür. *Panalgin* ilacı Reflü, Yüksek Tansiyon ve Depresyon için kullanıldı. Düzgüm derecesi 3’tür. *Minoset* ilacı sadece Migren hastalığına yazıldı. Bu şekilde ağrı kesici ilaçlar üzerinden Reflü, Yüksek Tansiyon, Migren, Depresyon, Prostat hastalıkları ve ilaçları arasında bağlantı olduğu görülmektedir. Bu hastalıklar ve ilaçlar diğerlerinden ayrı gruplanmış durumdadır. Diğerleriyle bir bağlantıları mevcut veritabanında yoktur. *Aspirin* ilacı Hemoroid ve Kolesterol hastalıklarına yazıldı. *Novalgin* ilacı ise sadece Hemoroid hastalığına yazıldı. Diğer hastalıklara ise herhangi bir ağrı kesici ilaç yazılmadı. Yani

yapılacağından bu kayıplar göz ardı edilebilir. Tek moda dönüşen bu ağlar üzerinde analizler yapılabilir. Bir önceki bölümde anlatılan *Ortak Komşu*, *Adamic Adar*, *Jaccard*, *Kaynak Tahsisli* yerel benzerlik algoritmaları kullanılarak bağlantı analizi yapıldı. Bu yöntemlerin seçilmesinin sebebi Liben-Nowell ve arkadaşları [36] ve Zhou ve arkadaşlarının [42] yerel benzerlik indislerini kullanarak gerçek ağlar üzerinde kıyaslamalar yapmaları ve kullanılan yöntemler içerisinde en iyi sonuçları bu yöntemlerin vermesidir.

Tek moda dönüştürülen ağlardan hastalık ağı Şekil 4.2'deki duruma dönüştü. Hastalık ağında Reflü ve Yüksek Tansiyon derecesi en yüksek olan düğümlerdir. Kolesterol ve Hemoroid'in sadece birbirleriyle bağlantısı vardır. Sivilce, Epilepsi ve Parkinson'un düğüm dereceleri 0 olduğu için diğer düğümlerle bağlantıları yoktur.



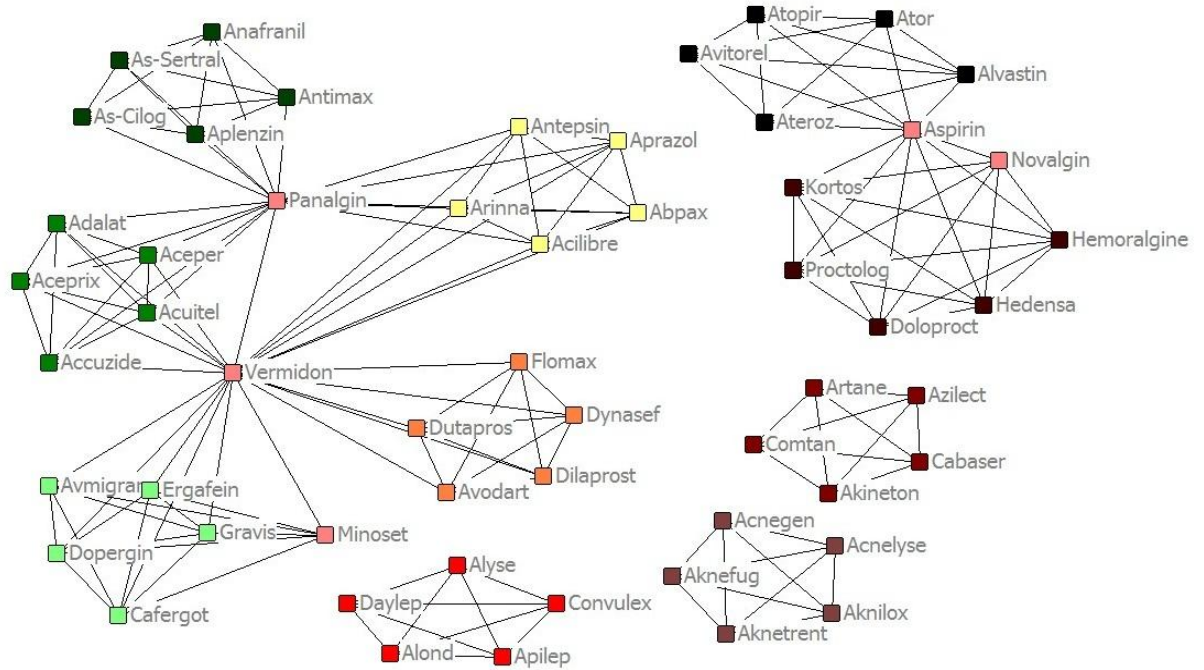
Şekil 4.2. Hastalık ağı

Tablo 4.2'de görüldüğü gibi, ağda Migren ve Depresyon ile Prostat ve Depresyon arasında doğrudan bağlantı olmadığı halde yapılan hesaplama göre bağlantı oluşma ihtimali oluştu. Migren ve Depresyon hastalıklarına aynı ilaçlar yazılmadı. Migren hastalığına *Vermidon*, Depresyon hastalığına *Panalgin* ilacı yazıldı. Bu iki ilacın ortak kullanıldığı hastalıklar olduğu için doğrudan olmasa da dolaylı oluşan bir ortaklıkları ortaya çıktı. Bu da ilerde beraber yazılma ihtimali ortaya çıkardı. Aynı durum Depresyon ve Prostat için de geçerlidir. Ortak komşuları Reflü ve Tansiyon üzerinden dolaylı bir bağlantı oluştu. Hesaplamalarda bu hastalık ilişkilerinin oluşma ihtimali aynıdır.

Tablo 4.2. Hastalık bağlantı tahmini değerleri

Hastalık İlişkisi	Ortak Komşu	AdamicAdar	Jaccard	KaynakTahsisli
Migren-Depresyon	[Reflü,Tansiyon]	1,44	0,67	0,50
Depresyon-Prostat	[Reflü,Tansiyon]	1,44	0,67	0,50

İlaç ağı Şekil 4.3'teki gibi 128 kenar ve 30 düğümden oluştu. Ortalama düğüm derecesi 8,5333'dir. Ağda alt ağlara geçiş ağrı kesici ilaçlar tarafından sağlandı. Birbiriyle ilişkide olmayan beş ayrı alt ağ oluştu.. Yoğunluğun en fazla olduğu yer düğüm dereceleri yüksek olan *Panalgin* ve *Vermidon* ilaçları üzerinden oluştu. Diğer bir alt ağda da ilişki *Aspirin* ilacı üzerinden oluştu. *Novalgin* ise sadece Hemoroid ilaçları ile beraber yazıldığı için ağda Hemoroid ilacı gibi rol aldı.



Şekil 4.3. İlaç ağı

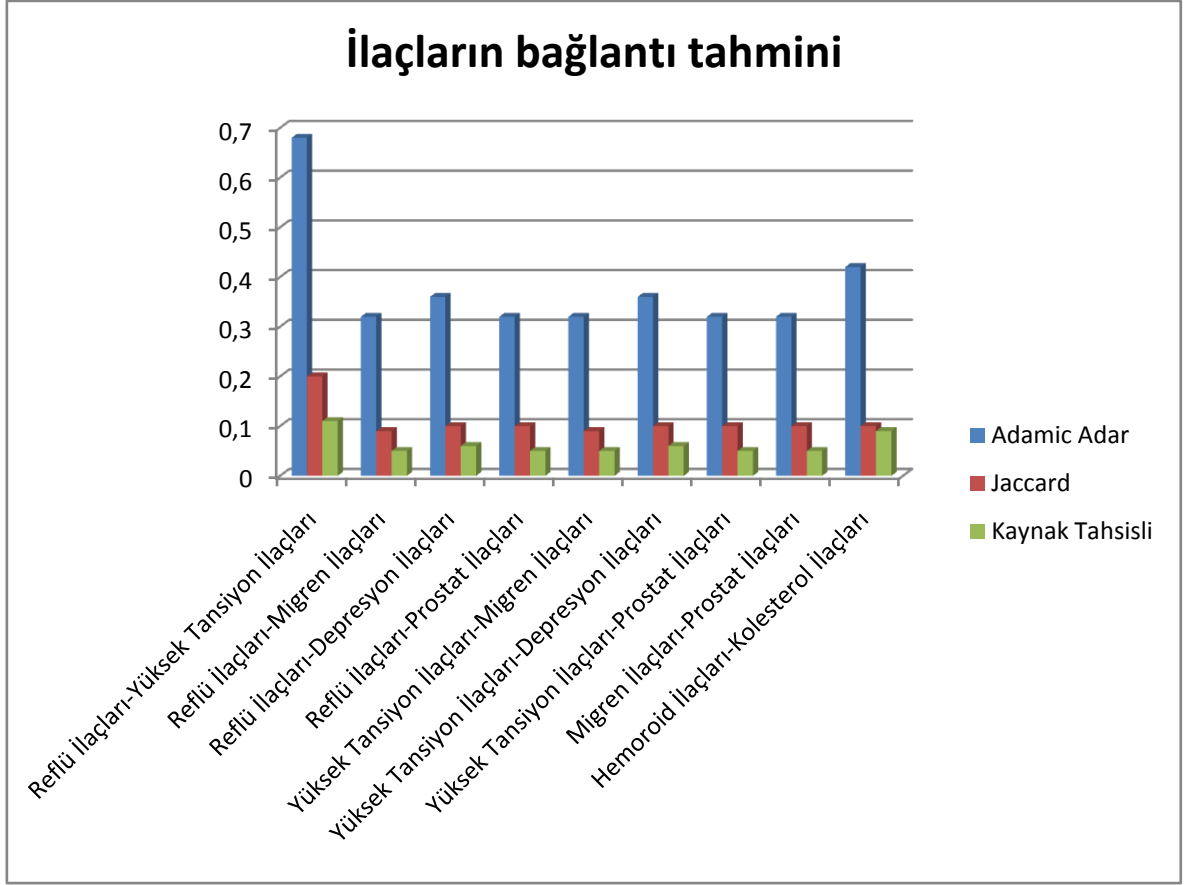
Oluşan veritabanında ağrı kesiciler dışında hiçbir hastalığın ilacının diğer hastalığa yazılmadığı Şekil 4.3'ten anlaşılmaktadır. Bağlantıları gözlemlemek için ağrı kesici ilaçlar

kullanıldı. İlaçlar beraber kullanılmamasına rağmen, ilaçların beraber yazılma ihtimali üzerine belirtilen yöntemlerle yapılan hesaplamalarda Tablo 4.3'teki değerler elde edildi.

Tablo 4.3. İlaç bağlantı tahmini değerleri

İlaç Bağlantıları	Adamic Adar	Jaccard	Kaynak Tahsisli
Reflü İlaçları-Yüksek Tansiyon İlaçları	0,68	0,20	0,11
Reflü İlaçları-Migren İlaçları	0,32	0,09	0,05
Reflü İlaçları-Depresyon İlaçları	0,36	0,10	0,06
Reflü İlaçları-Prostat İlaçları	0,32	0,10	0,05
Yüksek Tansiyon İlaçları-Migren İlaçları	0,32	0,09	0,05
Yüksek Tansiyon İlaçları-Depresyon İlaçları	0,36	0,10	0,06
Yüksek Tansiyon İlaçları-Prostat İlaçları	0,32	0,10	0,05
Migren İlaçları-Prostat İlaçları	0,32	0,10	0,05
Hemoroid İlaçları-Kolesterol İlaçları	0,42	0,10	0,09

Tablo 4.3'ten, Reflü ilaçları ile Yüksek Tansiyon ilaçlarının birbirine yazılma ihtimalinin daha yüksek olduğu oluşan değerlerden anlaşılmaktadır. Üç yöntemde de en yüksek değerler bu ilişkide ortaya çıkmaktadır. Diğer en yüksek oran Hemoroid ve Kolesterol ilaçları arasında oluşmaktadır. Jaccard indeksleme ile elde edilen değerler diğer iki yöntemle oransal olarak karşılaştırıldığında daha az değişkenlik göstermektedir. Şekil 4.4'ten bu daha net görülebilmektedir.



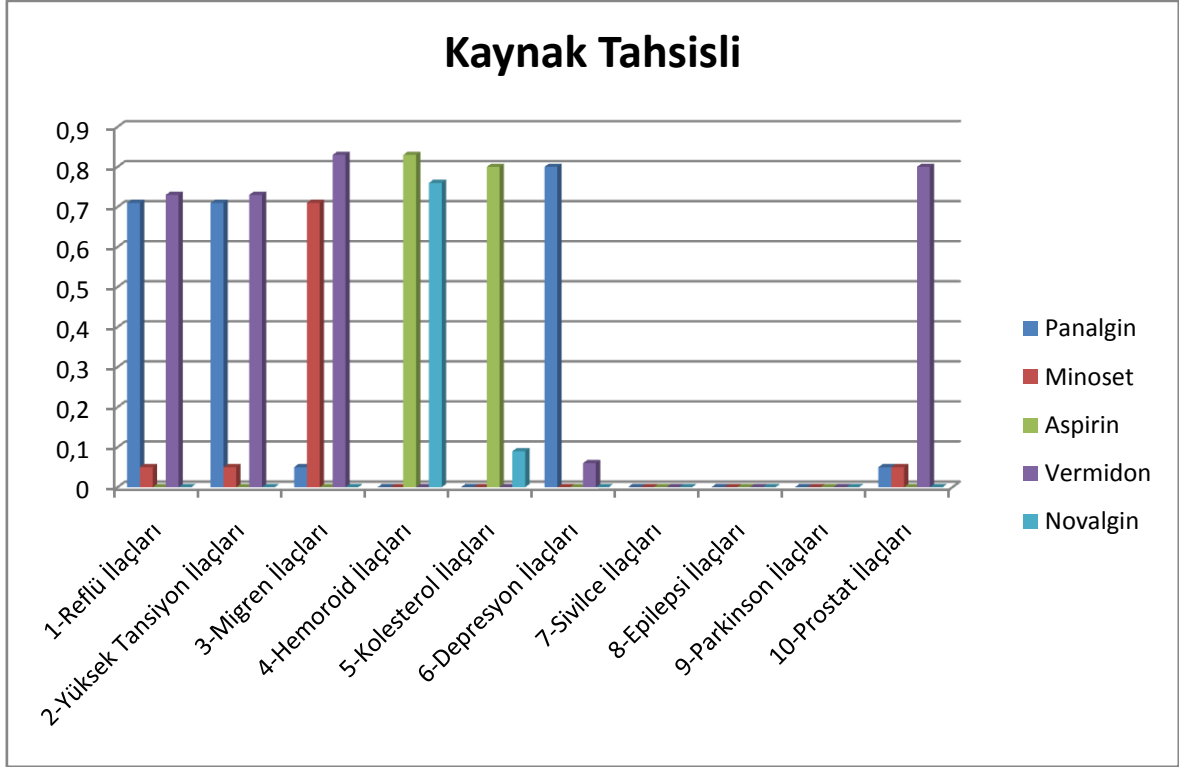
Şekil 4.4. İlaçların birlikte yazılma bağlantı tahmini grafiği

Veritabanında hastalıkların bazılarında ağrı kesici ilaçlar yazıldı. Zaten ilişki ağı bu ağrı kesici ilaçlar üzerinden oluşturuldu. Köprü görevi gören ağrı kesici ilaçlardır. Bu yüzden ağın kilit/köprü düğümleri bu ilaçlardır. Rastgele hastalıklara yazılan bu ilaçlar ağın yapısını ortaya çıkardı. Aşağıdaki tablolarda da hastalıklara yazılmadığı halde tahmin hesaplamalarında bazı hastalıklara yazılma ihtimali olan ağrı kesici ilaçlar koyu yazılarla belirtildi. Mesela en yüksek değeri veren *Novalgine* veritabanında Kolesterol hastalığına yazılan bir ilaç değildi. Ama sonuç tablolarına bakıldığında bu ilacın Kolesterol hastalığına yazılma ihtimali oluştu. Migren ve Depresyon arasında dolaylı bir bağlantı olmasına rağmen bu iki hastalığın ilaçlarının birbirine yazılma ihtimali gözüküyor.

Tablo 4.4. Ağrı kesici ilaçların bağlantı tahmini değerleri (Kaynak Tahsisli)

Kaynak Tahsisli					
	Panalgin	Minoset	Aspirin	Vermidon	Novalgin
1-Reflü İlaçları	0,71	0,05	0,00	0,73	0,00
2-Yüksek Tansiyon İlaçları	0,71	0,05	0,00	0,73	0,00
3-Migren İlaçları	0,05	0,71	0,00	0,83	0,00
4-Hemoroid İlaçları	0,00	0,00	0,83	0,00	0,76
5-Kolesterol İlaçları	0,00	0,00	0,80	0,00	0,09
6-Depresyon İlaçları	0,80	0,00	0,00	0,06	0,00
7-Sivilce İlaçları	0,00	0,00	0,00	0,00	0,00
8-Epilepsi İlaçları	0,00	0,00	0,00	0,00	0,00
9-Parkinson İlaçları	0,00	0,00	0,00	0,00	0,00
10-Prostat İlaçları	0,05	0,05	0,00	0,80	0,00

Kaynak Tahsisli İndeksleme ile yapılan hesaplamada Tablo 4.4'te oluşan değerler gösteriyor ki; Reflü hastalığına en fazla *Vermidon* ilacı yazılmıştır. *Minoset* ilacı Reflü hastalığına hiç yazılmadığı halde ilerde yazılma ihtimali ortaya çıkmıştır. *Aspirin* ve *Novalgin*'in Reflü hastalığına yazılma ihtimali ise yapılan hesaplamalarda yok gibi görünmektedir. Kolesterol ve Hemoroid hastalıkları arasında bağlantı *Aspirin* ilacı üzerinden oluşmuştu. *Novalgin* sadece Hemoroid hastalığına yazılmıştı. Tablo 4.4 ve Şekil 4.5'te 0,09 değeriyle *Novalgin* ilacının Kolesterol hastalığına yazılma ihtimali ortaya çıkmıştır.



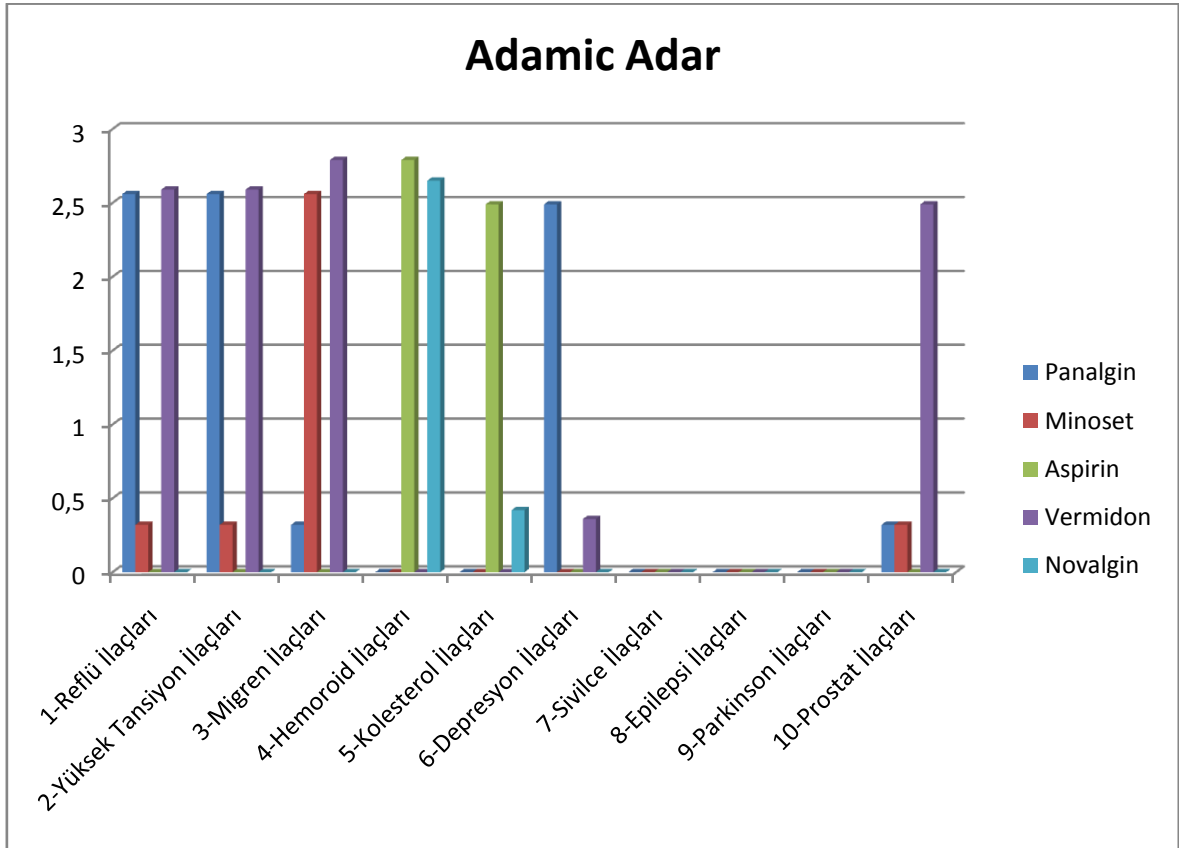
Şekil 4.5. Ağrı kesici ilaçların bağlantı tahmini sonuçları (Kaynak Tahsisli)

Şekil 4.5'ten anlaşıldığı üzere en yüksek değerleri Migren hastalığında *Vermidon*, Hemoroid hastalığında ise *Aspirin* elde etmiştir. Bu da en fazla bu hastalıklara yazılan ilaçların bunlar olduğunu göstermektedir.

Aşağıdaki Tablo 4.5 ve Tablo 4.6'da Adamic Adar ve Jaccard İndeksleme ile elde edilen değerler bulunmaktadır.

Tablo 4.5. Ağrı kesici ilaçların bağlantı tahmini değerleri (Adamic Adar)

Adamic Adar					
	Panalgin	Minoset	Aspirin	Vermidon	Novalgin
1-Reflü İlaçları	2,56	0,32	0,00	2,59	0,00
2-Yüksek Tansiyon İlaçları	2,56	0,32	0,00	2,59	0,00
3-Migren İlaçları	0,32	2,56	0,00	2,79	0,00
4-Hemoroid İlaçları	0,00	0,00	2,79	0,00	2,65
5-Kolesterol İlaçları	0,00	0,00	2,49	0,00	0,42
6-Depresyon İlaçları	2,49	0,00	0,00	0,36	0,00
7-Sivilce İlaçları	0,00	0,00	0,00	0,00	0,00
8-Epilepsi İlaçları	0,00	0,00	0,00	0,00	0,00
9-Parkinson İlaçları	0,00	0,00	0,00	0,00	0,00
10-Prostat İlaçları	0,32	0,32	0,00	2,49	0,00

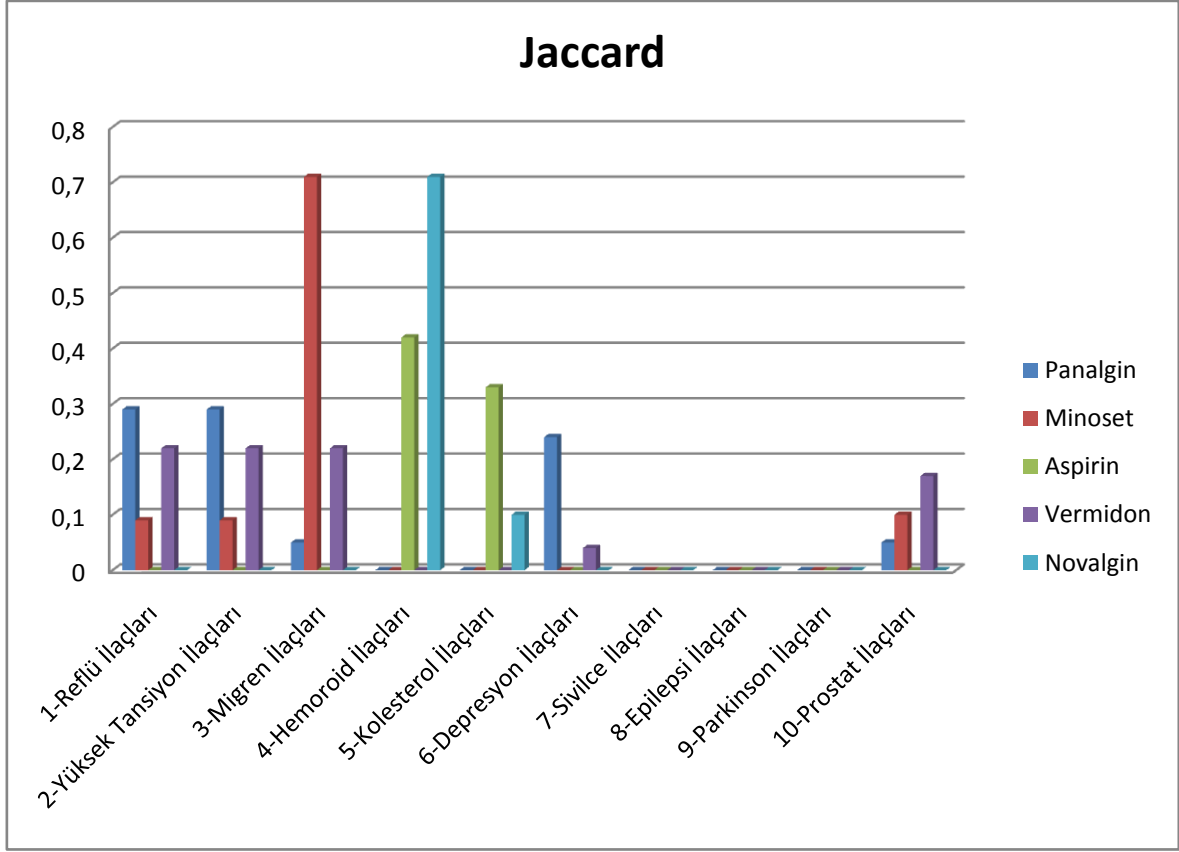


Şekil 4.6. Ağrı kesici ilaçların bağlantı tahmini sonuçları (Adamic Adar)

Tablo 4.6. Ağrı kesici ilaçların bağlantı tahmini değerleri (Jaccard)

Jaccard					
	Panalgin	Minoset	Aspirin	Vermidon	Novalgin
1-Reflü İlaçları	0,29	0,09	0,00	0,22	0,00
2-Yüksek Tansiyon İlaçları	0,29	0,09	0,00	0,22	0,00
3-Migren İlaçları	0,05	0,71	0,00	0,22	0,00
4-Hemoroid İlaçları	0,00	0,00	0,42	0,00	0,71
5-Kolesterol İlaçları	0,00	0,00	0,33	0,00	0,10
6-Depresyon İlaçları	0,24	0,00	0,00	0,04	0,00
7-Sivilce İlaçları	0,00	0,00	0,00	0,00	0,00
8-Epilepsi İlaçları	0,00	0,00	0,00	0,00	0,00
9-Parkinson İlaçları	0,00	0,00	0,00	0,00	0,00
10-Prostat İlaçları	0,05	0,10	0,00	0,17	0,00

Yapılan ölçümlerde Adamic Adar ve Kaynak Tahsisli yöntemlerinin verdiği ölçüm sonuçlarının birbirine yakın olduğu görülmektedir. Bazı ilaç türlerinde Jaccard indekslemenin sonuçlarının diğer iki yönteme göre bazı durumlarda değişkenlik gösterdiği belirlenmiştir. Mesela Migren hastalığına Kaynak Tahsisli ve Adamic Adar indekslemede *Minoset* ve *Vermidon* yazılması ihtimali birbirine yakın iken Jaccard'da bu ihtimal *Minoset* için yüksek olmasına rağmen *Vermidon* için daha düşüktür. Hemoroid hastalığına *Aspirin* ilacının yazılma ihtimali Kaynak Tahsisli ve Adamic Adar indekslemede daha yüksek olmasına rağmen Jaccard indekslemede *Novalgin*'e göre daha düşüktür. Reflü hastalığında Jaccard indekslemeye göre *Panalgin* ilacının yazılma ihtimali daha yüksek, diğerlerinde ise *Vermidon* çok fark olmasa da daha yüksek değerler vermektedir.



Şekil 4.7. Ağrı kesici ilaçların bağlantı tahmini sonuçları (Jaccard)

Bu değerlendirmelere bakıldığında Jaccard indekslemenin ağda daha farklı bir yaklaşımla sonuçlar verdiği söylenebilir. Jaccard indekslemenin bazı değerlerinin farklı olduğu görülebilmektedir. Şekil 4.5, Şekil 4.6 ve Şekil 4.7 üzerinden yapılan kıyaslamalarda açıkça görülmektedir. Ağdaki verilerimizin büyüklüğü arttıkça bu değerler farklılaşabilir. Kaynak Tahsisli ile Adamic Adar arasındaki benzerliklerde değişebilir.

Oluşturulan bu yapay veritabanı daha da büyütülebilir, ağın analizi daha da derinleştirilebilir. Ağın büyüklüğü arttıkça elde edilen sonuçların doğruluğunun tespiti zorlaşır. Fakat mevcut veritabanının analizi, ağ küçük olduğu için sonuçları değerlendirme açısından daha doğru gözlemler yapmayı sağlamıştır. Sonuçların sağlamlasının yapılması ve eksik noktalarının görülebilmesi açısından daha yararlı olmuştur.

Kullanılan yöntemler ortak komşular üzerinden hesaplamalar yapmaktadır. Sonuçlara bakıldığında ortak komşular yani iki düğüm arasındaki komşuların kesişimi ve birleşimi kullanılarak ağlar üzerinde önemli analizlerin yapılabileceği görülmektedir.

5.SONUÇ

Veri kayıtlarının tamamıyla sayısal ortamlara aktarıldığı günümüzde veriler somutlaşmış ve ölçülebilir hale gelmiştir. Tıp, akademik, sosyoloji, iş ve sayılmayan birçok alanda oluşan veri yığını üzerinde analizler yapılarak değerli bilgiler elde edilebilir. Günümüzde hastaneye giden hastaların bütün bilgileri kayıt altına alınmaktadır. Sayısal ortamlarda tutulan bu verilerin analiz edilmesi çok faydalı bilgilerin elde edilmesini sağlayacaktır. Sosyal ağlarda kullanılan analiz yöntemlerinden bağlantı tahmini yöntemleri bu verilerin analizi için kullanılabilir.

Bağlantı tahmininin kaynağı veri madenciliği ve sosyal ağlardır. Bu yüzden tezde başlangıçta bu konular üzerinde duruldu. Bu alanda yapılmış akademik çalışmalar incelendi. Oluşturulan veriler bir sosyal ağ verisiymiş gibi ele alındı.

Bu tez çalışmasında genel hatları ile farklı alanlarda kullanılan bu analiz yöntemlerinin hastalık-ilaç ağına uygulaması yapıldı. Hastalık ilaç ağlarının bağlantı tahmini yöntemleri kullanılarak analizi yapıldı. Özellikle yerel indeksleme yöntemlerinden Ortak Komşu, Adamic Adar, Jaccard, Kaynak Tahsisli indekslemenin başarılı sonuçlar vereceği düşünülerek, bu yöntemlerle analizler yapıldı. Düğümler arasındaki bağlantı ağı kesici ilaçlar köprü olarak kullanıldığı için onlar üzerinden sağlandı. Çünkü bir hastalığın ilacı diğerine yazılmadı. Sadece ağrı kesici ilaçlar diğer hastalıklara yazıldı. İlaçlar ve hastalıklar arasındaki ilişkilerin gerçekten işe yarar bilgiler barındırdığı ortaya çıktı. Nitekim hastanın hastalık ve bu hastalığa yazılan ilaçları üzerindeki verilerde yapılan analizler bize çok değerli bilgiler verdi. Mesela birbiriyle doğrudan ilişkili olmayan 2 ilacın beraber yazılma ihtimali ortaya çıktı. Hastalıklar arasında bağlantı ihtimali ortaya çıktı. Bağlantıların skor değerleri sıralanıp, hastalıklara hangi ilacın yazılma ihtimalinin yüksek olacağı belirlendi. Bu sonuçlar hastalıklara ilaç önerilerinin bu yöntemlerle sunulabileceğini gösterdi.

Jaccard indekslemenin diğer yöntemlere göre daha farklı bir yaklaşımla sonuçlar verdiği ortaya çıktı. Çünkü diğer yöntemler iki düğüm arasındaki ortak komşular ve komşuların dereceleriyle hesaplama yaparken, Jaccard ortak komşuların birleşimini de göz önüne aldı. Fakat genel olarak düğümlerin benzerlikleri birbirine yakın oranlarda ortaya çıktı.

Sosyal ağların barındırdıkları aktörlerin ilişkilerine ait oldukça faydalı bilgiler ihtiva ettiği hastalık ilaç ağı üzerinden bir daha gösterildi. Bu yapı ve ilişkilerin analiz edilmesi aracılığıyla yakınlıklar, benzerlikler, eğilimlere ulaşarak hastalık ilaç ağındaki ilişkiler hakkında çeşitli yorumlara ve tahminlere varıldı.

Bu tez çalışmasında komşuluk ilişkilerini kullanan yöntemler üzerinden hesaplamalar yapıldı. Hesaplamaların doğruluğu irdelendi. Uygulamada elde edilen bilgiler incelendiğinde yapılan tahminlerin gerçekçi sonuçlar verdiği görüldü.

Yapay veri yerine gerçek veriler kullanılarak uygulamanın kullanılabilirliği daha da artırılabilir. Sağlık uygulamalarında ve hastane otomasyonlarında bu algoritmalar kullanılarak kullanıcılara yönlendirmeler yaptırılabilir ve veriler üzerinde analizler yapılabilir. Nitekim bu uygulama ilerde daha büyük veriler üzerinde analiz yapmayı sağlayacak bir prototip niteliği taşımaktadır.

Bu alandaki çalışmalar içinde zor problemler barındırmaktadır. Verilerin büyüklüğüne göre karmaşıklık ve hesaplama zorluğu ortaya çıkmaktadır. Ama özellikle sosyal ağların geniş kesimler tarafından ilgi görmesi ve bu alandaki büyük verilerden kullanılabilecek işe yarar geniş analizlerin çıkma ihtimalinin olması sebebiyle son yıllarda üzerinde araştırmaların arttığı bir alandır.

Bağlantı tahmini açıklandığı üzere farklı alanlardaki birçok probleme çözüm sunacak ve birçok alanda uygulaması gerçekleştirilebilecek bir konudur. Özellikle sağlık alanı insanoğlunun ana hizmet alanlarından biridir. Bu alandaki hizmetler hataya sıfır tolerans göstermektedir. Bu alandaki hizmetler sürekli iyileştirilmeye çalışılmalıdır. Bu konuda yapılacak uygulamalarla geniş kesimleri ilgilendirecek çözümler üretilebilir. Sağlık çalışanlarının hizmet kalitesinin artırılmasına yardımcı olunabilir.

KAYNAKLAR

- [1] **Hwang, S.Y., Wylie, B.N., Wei, C.P. ve Liao, Y.F.**, 2010. "Coauthorship networks and academic literature recommendation", *Electronic Commerce Research and Applications*, 9, 323–334.
- [2] **Barabasi, A. L. ve Albert, R.**, 1999. "Emergence of scaling in random networks", *Science* 286, 509–512.
- [3] **Hasan, M.A., Chaoji, V., Salem, S. ve Zaki, M.J.**, 2006. "Link prediction using supervised learning", *SIAM 2006 Workshop on Link Analysis, Counterterrorism and Security*, Bethesda, MD.
- [4] **Folino, F. ve Pizzuti, C.**, 2010. "A comorbidity-based recommendation engine for disease prediction", *Proc of 23rd IEEE International Symposium on Computer-Based Medical Systems (CBMS 2010)*, 6-12.
- [5] **Folino, F. ve Pizzuti, C.**, 2012. "Link Prediction Approaches for Disease Networks", *ITBAM 2012*, 99-108.
- [6] **Alpaydın, E.**, 2000. "Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri", *Bilişim Semineri 2000*, 1-5.
- [7] **Özçakır, F. C. ve Çamurcu, A. Y.**, 2007. "Birliktelik Kuralı Yöntemi İçin Bir Veri Madenciliği Yazılımı Tasarımı Ve Uygulaması", *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 12, 21-37.
- [8] **Öğüt, S.**, 2009. "Veri Madenciliği Kavramı ve Gelişim Süreci", *Görsel İletişim Tasarımı Bölümü, İletişim Fakültesi, Yeditepe Üniversitesi, İstanbul*.
- [9] **Travers, J. ve Milgram, S.**, 1969. "An Experimental Study of the Small World Problem.", *Sociometry*, Vol. 32, No. 4, s. 425-443.
- [10] **Onat, F. ve Alıkhıç Ö.**, 2008. "Sosyal ağ sitelerinin reklam ve halkla ilişkiler ortamları olarak değerlendirilmesi", *Journal of Yasar University*, 3(9), 1111-1143.
- [11] **Özmen, Ş.**, 2009. "Ağ Teknolojisinde Yeni Ticaret Yolu: E-Ticaret", *İstanbul Bilgi Üniversitesi Yayınları, İstanbul*.
- [12] **Boyd, D.M. ve Ellison, N.B.**, 2007. "Social Network Sites: Definition, History and Scholarship", *Journal of Computer-Mediated Communication*, 13 (1), s. 210-230.

- [13] **Carminati, B., Ferrari, E. ve Perego, A.,** 2006. “Rule-Based Access Control for Social Networks”, Lecture Notes in Computer Science, 4278, s. 1734-1744.
- [14] **Pustylnick, I.,** 2011. “Patterns of Concealed Advertising in Social Network Websites”, Social Science Research Network.
- [15] **Kara, Y. ve Coşkun, A.,** 2012. “Sosyal Ağların Pazarlama Aracı Olarak Kullanımı: Türkiye’deki Hazır Giyim Firmaları Örneği”, Afyon Kocatepe Üniversitesi İİBF Dergisi, c.14, s. 2.
- [16] **Kuduğ, H.,** 2011. “Anlamsal Web Tabanlı Kurumsal İş Ağının Analizi”, Yüksek Lisans Tezi, Ege Üniversitesi Fen Bilimleri Enstitüsü, İzmir.
- [17] <http://techcrunch.com/2010/04/21/facebook-500-million-visitors-comscore/>
Facebook kullanımı 15.03.2013.
- [18] **Karcı, A. ve Boy, O.,** 2011. “Sosyal ağların web madenciliği teknikleri ile analizi ve ortak atıf analizi ile benzerlik tahmini”, Elektrik-Elektronik ve Bilgisayar Sempozyumu, Fırat Üniversitesi.
- [19] **Şahin, B.,** 2009. “Sosyal Ağ Analizi Nedir?”, ISREF.
- [20] **Gastner, M.T.,** 2011. “Networks: Theory and Applications”, Imperial College.
- [21] **Bondy, J.A., Murty, U.S.R.,** 2008, “Graph Theory”, Springer.
- [22] **Huisman, M., ve van Duijn, M.A.J.,** 2011. “The SAGE Handbook of Social Network Analysis”, s. 578-600.
- [23] **Huisman, M., ve van Duijn, M.A.J.,** 2004. “Software For Statistical Analysis Of Social Networks”, University of Groningen.
- [24] **Ting, I.H.,** 2008. “Web Mining Techniques for On-line Social Network Analysis”, IEEE, 696-700.
- [25] **Scott, J.,** 2000. “Social Network Analysis: A Handbook,” Sage Publications, 209s.
- [26] **Mika, P.,** 2007. “Social Networks and the Semantic Web, Semantic Web and Beyond Computing for Human Experience”, Jain,R. and Sheth, A.(Eds.), New York, Springer Science and Business Media, 234s.

- [27] **Newman, M.E.J.**, 2008. “Mathematics of Networks”, 12pp., The New Palgrave Dictionary of Economics, Durlauf, S.N. and Blume, L.E. (Eds.), Palgrave Macmillan.
- [28] **Schank, T. ve Wagner, D.**, 2005. “Approximating clustering-coefficient and transitivity”, Journal of Graph Algorithms and Applications, 9: 265–275.
- [29] **Hanneman, R.A. ve Riddle M.**, 2005. “Introduction to social network methods”, Riverside, CA: University of California, Riverside.
- [30] **Volkova, S.**, 2009. “Link Prediction in Social Networks”, KDD Lab Course Project.
- [31] **Getoor, L. ve Diehl C.P.**, 2005. “Link Mining: A Survey”, SIGKDD Explorations, 7(2), 3-12.
- [32] **Menon, A.K. ve Elkan, C.**, 2011. “Link Prediction via Matrix Factorization”, ECML/PKDD 2, 437-452.
- [33] **Shibata, N., Kajikawa, Y., ve Sakata, I.**, 2012. “Link Prediction in Citation Networks”, Journal of The American Society for Information Science and Technology, 63(1), 78-85.
- [34] **Seglen, P.O.**, 1994. “Casual Relationship Between Article Citedness and Journal Impact”, Journal of the American Society for Information Science and Technology, 45(1), 1–11.
- [35] **Vinkler, P. ve Davidson, G.S.**, 2002. “Dynamic changes in the chance for citedness”, Scientometrics, 54(3), 421–434.
- [36] **Liben-Nowell, D. ve Kleinberg, J.**, 2003. “The link prediction problem for social networks”, Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM), NewYork: ACM Press, s. 556–559.
- [37] **Popescul, A. ve Ungar, L.H.**, 2003. “Statistical relational learning for link prediction”, Proceedings of the Workshop on Learning Statistical Models from Relational Data, IJCAI-2003.
- [38] **Hasan, M.A., ve Zaki, M.J.**, 2011. “A Survey of Link Prediction in Social Networks”, Social Network Data Analytics, Springer, s. 243-275.

- [39] **Clauset, A, Moore, C. ve Newman, M.E.J.**, 2008. “Hierarchical structure and the prediction of missing links in Networks”, *Nature*, 453, 98–101.
- [40] **Murata, T. ve Moriyasu, S.**, 2008. “Link Prediction based on Structural Properties of Online Social Networks”, *New Generation Computing*, 26(3), 245–257.
- [41] **Caragea, D., Bahirwani, V., Aljandal, W. ve Hsu, W.H.**, 2009. “Ontology based link prediction in the Live Journal Social Network” *Proceedings of the Eighth Symposium on Abstraction, Reformulation, and Approximation(SARA2009)*.
- [42] **Lü, L., Jin, C.H. ve Zhou, T.**, 2009. “Similarity index based on local paths for link prediction of complex Networks”, *Physical Review E*, 80, 046122.
- [43] **D. A. Davis et al.**, 2008. “Predicting individual disease risk based on medical history”, In *Proc. of the ACM Int. Conf. on Information and Knowledge Management(CIKM'08)*, 769-778.
- [44] **Davis D. A., Chawla N. V., Christakis N. A. ve Barabasi A. L.**, 2010. “Time to CARE: a collaborative engine for practical disease prediction”, *Data Mining and Knowledge Discovery*, 20:388-415.
- [45] **Moody J.**, 2001. “Race, School Integration, and Friendship Segregation in America”, *American Journal of Sociology* 107, 679–716 .
- [46] **Borgatti, Stephen P., Mehra A., Brass D.J. ve Labianca G.**, 2009. “Network Analysis in the Social Sciences”, *Science* 323 (5916), 892–895.
- [47] **Gençer, M.**, 2013. “Sosyal Ağlar”, İstanbul Bilgi Üniversitesi.
- [48] **Galaskiewicz, Joseph, Wasserman S., Rauschenbach B., Bielefeld W., ve Mullaney P.**, 1985. “The Influence of Corporate Power, Social Status, and Market Position on Corporate Interlocks in a Regional Network”, *Social Forces* 64 (2), 403–431.
- [49] **Vartak, S.**, 2008. “A survey on Link Prediction”, State University of New York.
- [50] **Lü, L., Zhou, T.**, 2011. “Link Prediction in Complex Networks: A survey”, *Physica A* 390, 1150-1170.

- [51] Sağlık Bakanlığı, Birinci Basamakta Akılcı Reçete Yazımı, Hıfzıssıhha Mektebi Müdürlüğü, HMM Yayın No: HMM-2011-36, Ankara, 2011.
- [52] WHO, Promoting Rational Use of Medicines: Core Components - WHO Policy Perspectives on Medicines, No. 005, 2002.

ÖZGEÇMİŞ

Adı Soyadı : İhsan TUĞAL

Doğum Yeri ve Tarihi : Muş - 10.10.1980

EĞİTİM BİLGİLERİ

Fırat Üniversitesi-Bilgisayar Mühendisliği-Lisans (1997 - 2002)

Fırat Üniversitesi-Fen Bilimleri Enstitüsü-Bilgisayar Mühendisliği- Y.Lisans (2011-)

İŞ DENEYİMİ

Bitlis Eren Üniversitesi-Tatvan MYO-Öğretim Görevlisi (2011-Devam Ediyor)

Muş Alparslan Üniversitesi-Mühendis (2008 - 2011)

Muş Milli Eğitim Müdürlüğü-Mühendis (2002 - 2008)