

**YÖNSÜZ AĞIRLIKLİ HASTALIK AĞINDA
BAĞLANTI TAHMİNİ**

YÜKSEK LİSANS TEZİ

Müh. Serpil GÜL

**Anabilim Dalı: Bilgisayar Mühendisliğı
Programı: Kuramsal Temeller**

Danışman: Prof. Dr. Mehmet KAYA

AĞUSTOS-2014

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

YÖNSÜZ AĞIRLIKLİ HASTALİK AĞINDA BAĞLANTI TAHMİNİ

YÜKSEK LİSANS TEZİ

**Müh. Serpil GÜL
(121129106)**

Anabilim Dalı: Bilgisayar Mühendisliğı

Programı: Kuramsal Temeller

Danışman: Prof. Dr. Mehmet KAYA

Tezin Enstitüye Verildiğı Tarih: 14 Temmuz 2014

AĞUSTOS-2014

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

YÖNSÜZ AĞIRLIKLI HASTALIK AĞINDA BAĞLANTI TAHMİNİ

YÜKSEK LİSANS TEZİ

Müh. Serpil GÜL

Tezin Verildiği Tarih: 14 Temmuz 2014
Tezin Savunulduğu Tarih: 1 Ağustos 2014

Tez Danışmanı : Prof. Dr. Mehmet KAYA (F.Ü.)
Diğer Jüri Üyeleri : Doç. Dr. Mustafa TÜRK (F.Ü.)
Yrd. Doç. Dr. Taner TUNCER (F.Ü.)

AĞUSTOS-2014

ÖNSÖZ

Yüksek lisans eğitimim süresince bilgi ve deneyimleriyle beni yönlendiren, desteklerini esirgemeyen ve her konuda yardımcı olmaya çalışan, tecrübelerini paylaşan değerli danışman hocam Prof. Dr. Mehmet KAYA 'ya teşekkürü borç bilirim.

Serpil GÜL

ELAZIĞ-2014

İÇİNDEKİLER

	<u>Sayfa No</u>
ÖNSÖZ	I
İÇİNDEKİLER.....	II
ÖZET	III
SUMMARY	IV
ŞEKİLLER LİSTESİ	V
TABLO LİSTESİ.....	VI
1. GİRİŞ	1
1.1. Tezin Amacı.....	3
1.2. Tezin Yapısı.....	4
2. SOSYAL AĞLAR VE ANALİZİ	5
2.1. Çizge Teorisi.....	12
3. BAĞLANTI TAHMİNİ.....	16
3.1. Bağlantı Tahmini Problemi.....	16
3.2. Yapılan Bazı Çalışmalar	18
3.3. İlişki ve Ağ Türleri	19
3.4. Ağırlıklı ve Ağırlıksız Ağlarda Bağlantı Tahmini Yöntemleri.....	20
3.5. Benzerlik Ölçütü Yöntemleri.....	23
4. HASTALIK AĞININ OLUŞTURULMASI VE ÖNERİLEN YÖNTEM ...	27
4.1. Giriş	27
4.2. Verinin Hazırlanması.....	27
4.3. Hasta Tanılarının ICD-10-Tr ile Kodlanması.....	30
4.4. Hastalık Ağının Oluşturulması	31
4.5. Komşuluk Düğümlerine Dayalı Bağlantı Tahmini Ölçütleri.....	33
4.6. Geliştirilen Yöntem	33
5. UYGULAMA SONUÇLARI	37
5.1. Oluşturulan Ağın Yapısal Olarak İncelenmesi	37
5.2. Bağlantı Tahmini Yönteminin Değerlendirilmesi	38
6. SONUÇ	43
KAYNAKÇA.....	45
ÖZGEÇMİŞ	48

ÖZET

Bağlantı tahminiyle 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.

Uygulamada Fırat Üniversitesi Hastanesinde Hemogram testi yaptıran hasta verileri kullanarak “*Yönsüz Ağırlıklı Hastalık Ağı*” oluşturuldu. Hastalık ağında, her bir düğüm birer hastalığı ilişkiler ise hastalıklar arası ilişkiyi temsil etmektedir. Sosyal ağ analizi ve bağlantı tahmini metotları kullanılarak yeni bir bağlantı tahmini yöntemi geliştirilmiştir. Geliştirilen bu yöntemle, belirli şikâyetlerle gelen hastaların ne gibi hastalık riskleri taşıdığını tespit edilerek proaktif bir yaklaşımla hastalık riski tahmini yapılmaktadır.

Bu yöntemlerin kullanılabilirliği, bağlantı tahminin bu ağlara uygulanabilirliği gösterildi.

Anahtar Kelimeler: Hastalık ağları, Bağlantı tahmini, Sosyal Ağ Analizi, Benzerlik Ölçütü Algoritmaları, Veri Madenciliği

SUMMARY

LINK PREDICTION IN UNDIRECTED WEIGHTED DISEASE NETWORK

Through the social network analysis we can get new data from various data and is intended to generate future predictions in the network structure. 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.

In this application, “Undirected Weighted Disease Network” is created with using patent information who completed blood test in Firat University Hospital. In Disease Network, each node is disease and each relation is relationship between disease. We developed an new link prediction approach with using social network analysis methods and link prediction methods. Through our approach, detecting what kind of diseases risks of patients have coming the hospital with specific complaints.

The availability of these methods and applicability of link prediction to these networks is shown.

Key Words: Disease Network, Link Prediction, Social Network Analysis, Proximity Metrics Algorithm, Data Mining

ŞEKİLLER LİSTESİ

	<u>Sayfa No</u>
Şekil 1.1	LinkedIn Arkadaşlık Sosyal Ağ Yapısı..... 2
Şekil 2.1.	Yeni nesil web 9
Şekil 2.2.	Sosyal ağ sınıfları 10
Şekil 2.3.	Düğümmler arasında oluşan ilişkiler..... 12
Şekil 2.4.	Königsberg'in 2 adası ve 7 köprüsü 13
Şekil 2.5.	Düğümmler, kenarlar ve kenarların düğüm kümeleri 14
Şekil 2.6.	Yönlü ve yönsüz çizgeler 15
Şekil 2.7.	Sosyal ağın çizge yapısı 15
Şekil 3.1.	Sosyal ağın zamansal değişimi 18
Şekil 3.2.	Amerika'daki bir okuldaki arkadaşlık ağı. Renkler etnik kökenlerini göstermektedir. 20
Şekil 3.3.	Bağlantı tahminine farklı yaklaşımlar 21
Şekil 3.4.	Ağ yapısındaki ilişkilerin zamansal değişimi..... 22
Şekil 4.1.	Veri Madenciliği 28
Şekil 4.2.	Yönsüz Ağırlıklı Hastalık Ağından ilk 22 Hastalık 32
Şekil 4.3.	Bağlantı Tahmini Hesabında kullanılacak üç ağ: (a) <i>G</i> , (b) <i>Geğitim</i> , ve (c) <i>Gyeni</i> 35

TABLO LİSTESİ

	<u>Sayfa No</u>
Tablo 2.1. Farklı bilim dallarında ağların farklı kelimelerle ifadesi.	13
Tablo 3.1. Sosyal ağ araştırmalarında karşılaşılan bağ türleri ve veri seti özellikleri	20
Tablo 4.1. Fırat Üniversitesi Hastanesi Hemogram Testi yaptıran Hasta Kayıt Verileri	29
Tablo 4.2. ICD-10-Tr sınıflama ölçütüyle kodlama aşamasından sonra oluşan veritabanı yapısı	30
Tablo 4.3. Hastalıklar arası ilişkiler.....	31
Tablo 4.4. Ağırlıklı Hastalık Ağı ilk 22 hastalık sosyomatrisi.....	32
Tablo 4.5. Herbir eksik bağlantı düğümünün Jaccard skoru	36
Tablo 5.1. Hasta – Hastalık Ağı’nda ağın en merkezi düğümleri	37
Tablo 5.2. Hasta – Hastalık Ağı’nda ağırlık derecesi en yüksek 10 bağlantı.....	38
Tablo 5.3. Jaccard Bağlantı Tahmini Ölçütü Top-10 Listesi	40
Tablo 5.4. Ortak Komşu Bağlantı Tahmini Ölçütü Top-10 Listesi.....	40
Tablo 5.5. Tercih Bağlantılı Bağlantı Tahmini Ölçütü Top-10 Listesi	40
Tablo 5.6. Adamic-Adar Bağlantı Tahmini Ölçütü Top-10 Listesi	41
Tablo 5.7. Geliştirilen yöntemin Başarı Oranları	41

1. GİRİŞ

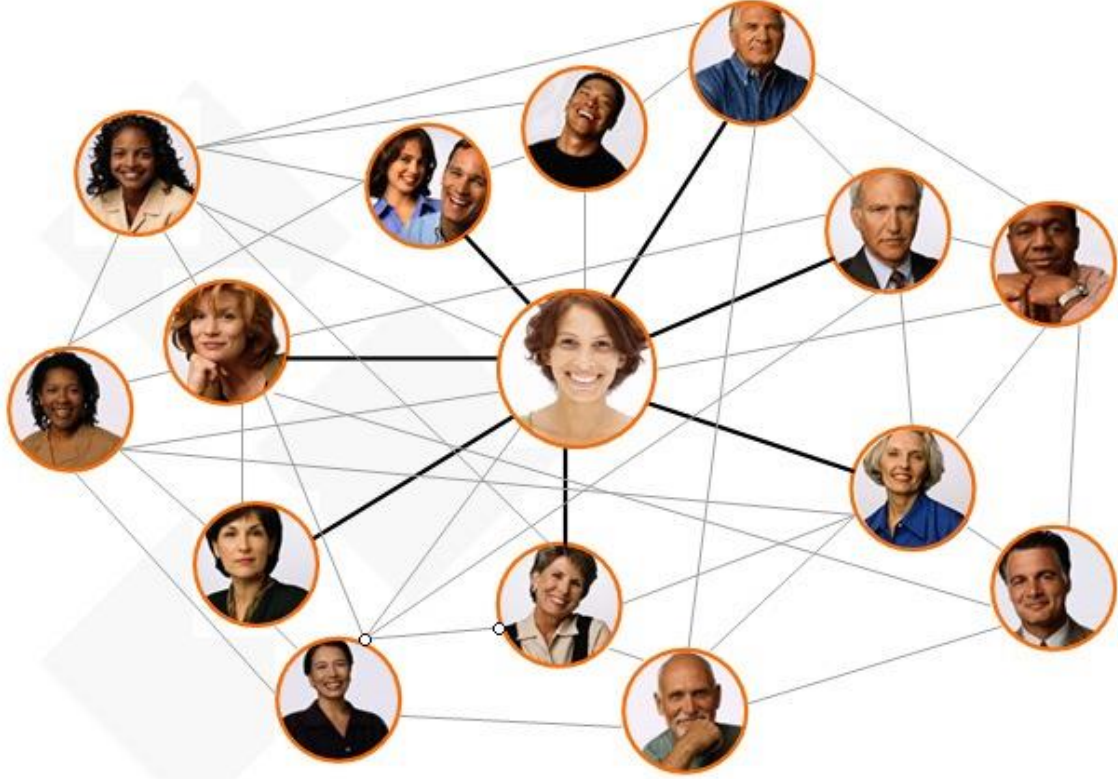
Son yıllarda gelişen internet ve iletişim teknolojilerinin toplumsal hayatı derinden etkilediği görülmektedir. İnsanların kolay ve hızlı bir şekilde internet hizmetinden yararlanmasıyla birlikte, internet ortamında geçirilen süreler de giderek artmaya başlamıştır. Özellikle sosyal web kavramının gelişi ile birlikte, web üzerinde bireyler arasındaki etkileşimi arttıran, sanal ortamda sosyal hayattan esintiler sunan uygulamalar oldukça popüler hale geldi. Bu uygulamalar içerisinde beklenmedik bir şekilde hızla gelişen ve kısa süre içerisinde milyonlarca insan tarafından kullanılmaya başlayan sosyal ağlar, insanlığı sanal ortamdaki yeni bir sosyal yaşam biçimine doğru sürüklemektedir.

Sosyal ağlar, bireyler (aktörler) ve bireyler arasındaki ilişkilerden oluşan yapılardır[] . Bireyler insanlar, organizasyonlar, canlılar, bilgiler olabilir. İlişkiler akrabalık, arkadaşlık, beğeni, iş birlikteliği, fonksiyonellik, kaynak transferi ya da bilgi akışı, satın alma vs. şeklinde olabilir. Sosyal ağların gelişimiyle birlikte, ağdaki düğümler(elemanlar) ve ilişkiler ile ilgili bir takım bilgileri ortaya çıkarmak amacıyla yapılan analiz çalışmaları da yapılmaya başlanmıştır. Sosyal ağ analizi (SAA) olarak adlandırılan bu çalışmalarda elde edilecek birtakım ölçütler sayesinde ağın genelinin yapısı ile ilgili bilgi sahibi olunabileceği gibi ağdaki herhangi bir eleman ve bu elemanın konumu ve ağ için önemi hakkında bilgiler elde edebilmekte mümkündür. Çünkü SAA ağdaki aktörler arası ilişkileri ve aktörlerin davranışlarını inceler. Sosyal ağların yapısını daha iyi anlamayı, sosyal ağlarda saklı olan bilgileri ortaya çıkarmayı hedefler.

Sosyal ağlarda görülen bu hızlı büyüme ve gelişme sayesinde ağ yapılarının çok çeşitli amaçlar için kullanımını da cazip hale getirmiştir. Sadece sosyal paylaşım siteleriyle sınılı kalmayıp; eğitim, sağlık, ekonomi, vs. alanlarda da uygulanmaya başlamış; biyologların, matematikçilerin, bilgisayar bilimcilerin, ekonomistlerin, sosyologların ve daha birçok bilim adamların ilgi alanına girmiştir.

Şekil 1.1 'de LinkedIn sosyal paylaşım sitesinden alınan küçük bir sosyal ağ yapısı gösterilmektedir. Bu ağ yapısına bakılarak, bir bireye erişmenin o bireyle doğrudan ilişkiye sahip olmayı gerektirmediği görülür. Bu tarz sitelerde internet kullanıcılarının arkadaşlarının, ilgi alanlarının, alışverişlerinin, internet sayfalarındaki gezintilerinin analizi, ağdaki aktörler arasında olası bağlantı tahmini [1,2,3], kullanıcıların

davranışlarından örüntü çıkarımı [4], topluluk keşfi [5,6,7], terörist ağının analizi ile terörist gruplarına karşı önlem alma SAA'nin uygulama alanlarından bazılarıdır.



Şekil 1.1 LinkedIn Arkadaşlık Sosyal Ağ Yapısı

Şekil 1.1' deki gibi ağlar sayesinde dünyanın öbür ucunda hiç tanımadığımızın tanıdığı zinciriyle ulaşabilmemiz mümkündür. Bu yaklaşımla hareket eden ünlü sosyopsikolog Stanley Milgram 1967'te Amerika'da dünya küçüktür fenomenini ortaya atmıştır. Bunun için şöyle bir deney yapılmıştır; Amerika'da bir grup insandan bir mesajı belirlenmiş bir kişiye yakınları arkadaşları vasıtasıyla ulaştırmaları istenmiştir. Deney sonucunda ortalama beş kişi ile bu mesajın bu insana ulaştığı görülmüştür. Kolombiya Üniversitesi'ndeki araştırmacılar bu deneyi ilerletmiş, dünya üzerinde herhangi iki kişinin internet ya da telefon üzerinden iletişime geçmesi ortalama 5-7 aracı ile gerçekleşmiştir [8]. Dünyadaki insanların ya da başka canlıların dünyayı saran bir örümcek ağı gibi birbirlerine bağlı olduğu yapıları incelemek analiz etmek birçok noktada ufuk açıcı şeyler karşımıza çıkarmakta ve çıkaracaktır.

Örneğin, ülkeler arası ekonomik ilişkiler bir sosyal ağda gösterilebilir. Ülkeler arası ticaretler, yatırımlar bu ağda analiz edilerek ülkelerin ekonomik durumları başka ülkelere

etkileri daha iyi anlaşılabilir. Daha da önemlisi sosyal ağlar sağlık sektöründe kullanılarak milyonlarca insanın hayatında çok önemli kolaylık sağlayacak uygulamalar geliştirebilir.

1.1. Tezin Amacı

Tıbbi bakım ve araştırma insanlar için tam anlamıyla bilimin en önemli parçasıdır. Hiç birimizin fiziksel rahatsızlıklara ve biyolojik bozulmaya karşı koyacak bağışıklık sistemi yoktur. Dünyadaki yıllık sağlık bakım masrafları inanılmaz boyuttadır. 2001 yılında yapılan bir araştırmaya göre yılda 880,5 milyon hastanın doktor muayenelerini ziyaret ettiği ve doktor başına 3,1 milyon hastanın düştüğü tahmin edilmektedir [2]. Bu büyük artış bilim insanlarını, hastalık risklerini önceden araştırıp tahmin ederek önlemler almaya teşvik etmektedir. Bu alanda yapılan tahminlerin reaktiften ziyade proaktif olması gerekir. Bu yüzden hastanın hastalık geçmişi ve benzer hastaların hastalık geçmişleri kullanılarak hastalık riskleri tespit edildikten sonra “*Hastalık oluşmadan belirtileri tedavi edilir.*” yaklaşımı uygulanmalıdır [1].

Amaç, belirli şikâyetlerle gelen hastaların ne gibi hastalık riskleri taşıdığını tespit ederek proaktif bir yaklaşımla hastalık riski tahmini yapmaktır. *Yüksek lisans tez çalışmasında, hastane veya doktorlara herhangi bir öneri platformu oluşturma amacı güdülmemiştir.* Yapılan akademik çalışmanın sonuçlarını test etmek için gerçek veri kullanımına ihtiyaç duyulmuştur. Çalışmanın herhangi bir yerinde hastaya ait tanımlar hariç diğer bilgiler kesinlikle kullanılmamıştır. Tasarlanan ağ üzerinde tanımların hangi hastaya ait olduğu gizlenmiştir. Bu sistemi geliştirirken veri madenciliğini, sosyal ağ modelleme tekniklerini, sosyal ağ analizi, ve bağlantı tahmini algoritmaları kullanılmıştır.

Gerçekleştirilen bu yüksek lisans tezinde öncelikle bir hastalık ağı oluşturulacaktır. Bu ağda her bir düğüm bir hastalığı gösterirken düğümler arasındaki bağlantı ise iki hastalığa zamanla yakalanan hastaları ifade eder. En az iki kez hastaneye gelen ve en az iki farklı teşhis konulmuş 1000’den fazla hastadan oluşturulacak hastalık ağının bir kısmı eğitime verisi diğer kısmı ise test verisi olarak kullanıldı. Bağlantı tahmini algoritmalarından yüksek performanslı sonuçlar almak için çalışmada gerçek verilere ihtiyaç duyuldu. Bu hususta geliştirilen yöntemin doğruluğunu tespit etmek için gerçek veriler kullanılmıştır. Çalışmada “*2013 Yılı Fırat Üniversitesi Hastanesi Hemogram Testi Yapılan Hasta Kayıt Verileri*” kullanılmıştır. Bu veriler, projenin etik olarak uygunluğu Fırat Üniversitesi

Girişimsel Olmayan Araştırmalar Etik Kurulu tarafından değerlendirildikten sonra Fırat Üniversitesi Biyokimya Laboratuvar'ından alınmıştır.

Çalışmanın ilham kaynağı Folino ve Pizzuti'nin 2012 yılında yaptıkları bir çalışmadır [5]. O çalışmada her bir düğüm bir hastalığı, bağlantı ise bir kişinin ilgili hastalıklara aynı anda yakalanma durumunu temsil eder. Bağlantı tahmini ile aynı anda hangi hastalıkların bir arada oluşabileceği belirlenmiştir. Veri olarak da İtalya'nın güneyindeki bir kasabadan elde edilen hastalık bilgileri kullanılmıştır. Çalışmanın diğer bir ilham kaynağı ise Davis ve arkadaşlarının [1,2] 2008 ve 2010 yıllarında hastanın geçmiş hastalık bilgilerinden hastalık tahmini yapmak için yöntemler önerdikleri çalışmalarıdır. Yüksek lisans tez çalışmasında amaç, farklı bir bağlantı tahmini yöntemiyle ilişkileri ağırlıklı hastalık ağından yeni bağlantıların tahmin edilmesidir.

1.2. Tezin Yapısı

Tez çalışmasının ikinci bölümünde; sosyal ağların tanımı, kapsamı ve modellenmesi anlatılmıştır. Üçüncü bölümde; bağlantı tahmini problemi tanımı, bu alanda yapılan çalışmalar, bağlantı tahmini yöntemleri yer almaktadır. Dördüncü bölümde; problemin çözümü için önerilen yöntem tanıtılmış, uygulama adımlarından bahsedilmiştir. Beşinci bölüm; tez çalışması süresince elde edilen sonuçları ve ileriye dönük düşünceleri içermektedir.

2. SOSYAL AĞLAR VE ANALİZİ

Günlük hayatta pek farkında olmasakta evren üzerinde dört bir yanımız çeşitli ağ yapılarıyla çevrilidir. Örneğin, internet ağı, canlılar ve çevre arasındaki ekolojik ağlar, haberleşme ağları, dağıtım ağları, toplumsal ve kültürel ağ yapıları, karayolu, demiryolu ve havayolu ağları. Bu örnekleri daha da çoğaltmak mümkündür [18]. Ağ kavramı kelime anlamı olarak “birçok nokta ile bunlar arasındaki bağlantılarla gösterilebilen bir dizgeye ilişkin yapı” olarak tanımlanmaktadır [18]. Sosyal bir ağ ise, düğümler olarak adlandırılan aktörlerden veya elemanlardan oluşan ve bu düğümleri kendi aralarındaki çeşitli ilişki ve etkileşimlere bağlı olarak birbirine bağlayan bir yapı olarak tanımlanabilir. Birbirleri ile etkileşim içerisinde bulunan bu varlıkların sosyal ağ yapısı içerisinde gerçekleştirdikleri her türlü etkileşimin detaylı olarak incelenmesi ve analiz edilmesi sonucunda söz konusu yapıyla ve içerdiği düğümlerle ilgili olarak ilk bakışta fark edilemeyen çok kıymetli bilgiler elde edilebilmektedir.

Sosyal ağ kavramı internetin başlangıcından beri var olan bir kavramdır. 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 [6].

Günümüzde iletişim teknolojilerinde ve özellikle internet ortamında yaşanan her olay somut veriler olarak depolanmaktadır. 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. Dijital ortamda sağlanan somut veri kaynağının çok büyük boyutlarda olması çeşitli problemlere yol açmaktadır. Bu verilerin temizlenmesi, işlenmesi ve yararlı verilerin elde edilmesi oldukça karmaşık işlemlerdir.

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 bilgisayar ortamlarının bu denli yoğun kullanılmasıyla; veri, enformasyon, bilgi vb. birçok kavrama aşına olmuş durumdayız. Çevremizdeki birçok

alandaki kullanılan bilişim teknolojisi ve bunun yanında kişisel bilgisayar kullanımıyla birlikte birçok bilgisayar kavramı da hayatımıza girmiştir. Bunlardan en önemlileri “veri” ve “bilgi” kavramlarıdır. Günlük hayatımızda sıkça kullanılır hale geldiğimiz bu kelimeler yalın halde kullanıldığında gerçek anlamlarından uzaklaşmaktadırlar. Veri, bilgi sistemleri, işletme içinde ki insanlar, nesneler ve bunlar arasında ki ilişkiler hakkında bilgiler içermektedir. Veri kavramını dünyanın betimlemesi olarak düşünebiliriz. Duyu organlarımızla verileri algılar ve beynimizle işleriz. Bilgi ile sözü edilen veri, işlenerek karar verici, anlamlı ve kullanışlı hale gelir. İnsanlar bilgiyi oluşturmak için uzunca bir süre veriyi kullandılar.

Bilginin bu denli değerli olması, bilişim teknolojilerinin gelişmesine ön ayak olmuştur. Bu gelişmelerle birlikte bilgisayarlar bilgi üretiminde ve denetiminde kayda değer bir öneme sahip olmuştur. 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].

Büyük miktardaki veriler içerisinde önemli olanlarını bulup çıkarmaya Veri Madenciliği denir. Veriler üzerinde çözümlemeler yapmak amacıyla ve veriyi çözümleyip bilgiye ulaşabilmek için veri madenciliği yöntemi ortaya çıkmıştır. Veri madenciliği bir sorgulama işlemi veya istatistik programlarıyla yapılmış bir çalışma değildir. Veri madenciliği milyarlarca veri ve çok fazla değişken ile ilgilenir. Teknolojik gelişmeler dünyada gerçekleşen birçok işlemin elektronik olarak kayıt altına alınmasını, bu kayıtların kolayca saklanabilmesini ve gerektiğinde erişilebilmesini hem kolaylaştırıyor, hem de bu işlemlerin her geçen gün daha ucuza mal edilmesini sağlar. Ancak, ilişkisel veri tabanlarında saklanan birçok veriden kararlar için anlamlı çıkarımlar yapabilmek bu verilerin bilinçli uzmanlarca analiz edilmesini gerektirir. Veri sayısı çok olduğu için bazı özel analiz algoritmaları geliştirilmiştir. 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.

Yaşadığımız bu zamanda dijital iletişim teknolojilerinin bu denli 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 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 (bileşenlerine ayırma, kümeleme, ilişkilerin belirlenmesi), elektronik ticaret ve çevrimiçi reklamcılık (müşteri profilinin çıkarılması ve eğilim analizi, kişiye özel reklamcılık ve teklif sunma), fiziki yapıların analizi (ulaşım, tesisat, altyapı) ve büyük veri kümelerinin

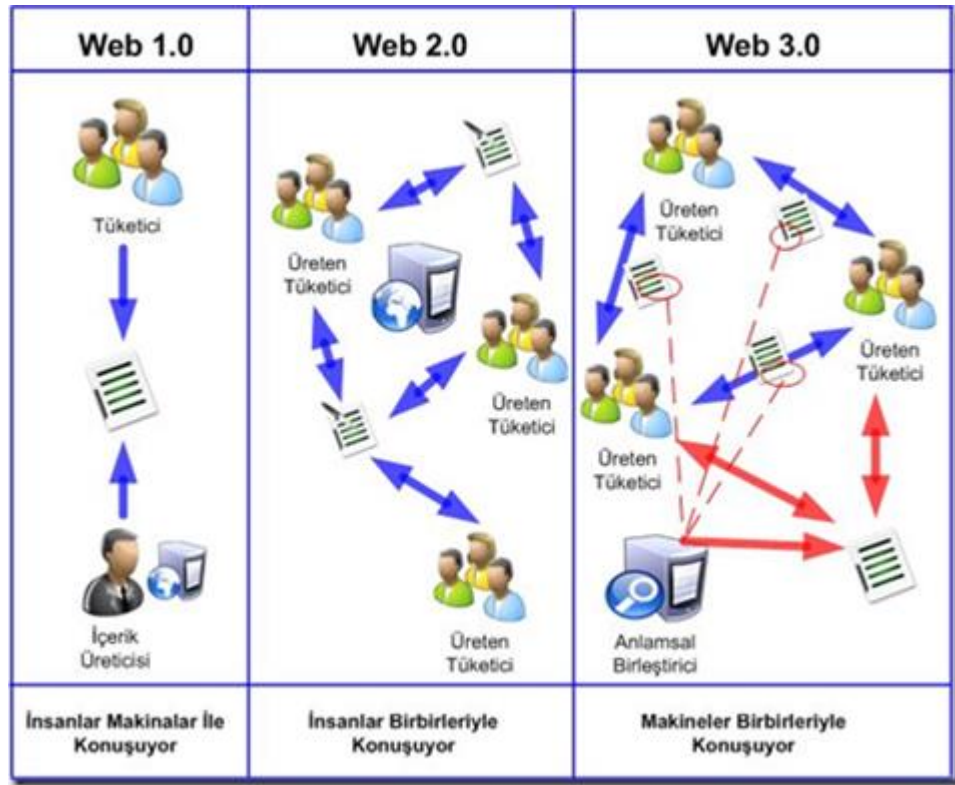
analizi (medya takibi, akademik yayın analizi, genetik arařtırmalar) gelmektedir. Sosyal ađ siteleri, sosyal bir evre oluřturmak amacıyla kurulan, byk kitlelerin birbirleriyle yođun olarak iletiřime ve etkileřime geebildikleri elektronik ortamlardır. zmen [11], sosyal ađ sitelerini bireylerin, kendilerine zel bir alan oluřturarak sistemin diđer yeleriyle bir araya gelerek dijital ortamda paylařım ve etkileřim iine girmelerine imkn sađlayan sanal ortamlar olarak tanımlamıřtır. Carminati ve arkadaşları [12] web tabanlı sosyal ađları; web kullanıcılarının kaynak paylařımına ve iř, eđence, flrt gibi amalarla diđer kullanıcılarla arkadaşlık kurmalarına olanak sađlayan evrimii topluluklar olarak tanımlamıřtır. Boyd ve Ellison [13], sosyal ađları bireylerin sınırlı bir sistem ierisinde halka aık veya yarı aık bir profil oluřturmalarına, bađlantılı oldukları diđer kullanıcıların bir listesini beyan etmelerine, kendilerinin ve sistemdeki diđer kullanıcıların bađlantı listelerini grntlemelerine ve geiř yapmalarına izin veren web tabanlı servisler olarak tanımlamıřtır.

Var olan birok sosyal ađ sitesi iřleyiřleri bakımından genel olarak iki grup altında incelenmektedir [14]. Birinci grup; her kullanıcının kendine ait sayfalarının bulunduđu kullanıcı tabanlı sosyal ađ siteleridir. Bu sitelerde zel konu bařlıđı olmaksızın, sistemdeki tm kullanıcılara hitap ederler. Facebook, Myspace gibi siteler rnek olarak verilebilir. İkinci grup; İnsanları belirli bir konu, hobi veya dřnce etrafında toplayan grup tabanlı sosyal ađ siteleridir. Belli bir ilgi veya iliřki esas alınarak oluřturulurlar. LinkedIn, Academia ve Ravelry bu grup sosyal ađlara rnek gsterilebilir [15].

Yařamımızın nemli bir parası haline gelen internet ortamında her gn milyonlarca insan iletiřim kurmaktadır. Kiřilerin tanıdıđı veya tanımadıđı kiřilerle etkileřime geebildiđi, ierik retebildiđi, kendi dnyasını kurduđu ortamlardır. İlk bařlarda web teknolojisi bu dinamikliđe sahip deđildi. Web teknolojisi geliřim ařamaları řekil 2.1’de grldđ gibidir. zellikle Web 2.0’ dan sonra hızlı bir bymeye giren sosyal ađlar insanların yařamını nemli dzeyde etkilemeye bařlamıřtır. Yař sınırı olmaksızın bugn neredeyse herkes bu sitelerden haberdardır. İnsanların yz yze iletiřim kurmasını gerektirmeyen bu yapılar, farklı cođrafyalardan, farklı kltrlerden olan bireylerin birbirleriyle kolayca etkileřim kurmalarını sađlamaktadır. Bu sayede sosyalliđi azalan gnmz insanları, kendilerini rahata ifade edebildikleri bu siteleri kullanarak daha aktif ve sosyal bir yařam srmektedirler.

Sosyal bir devrim olarak tanımlanan Web 2.0' dan önce sosyal ağlar kullanılmaktaydı. 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.

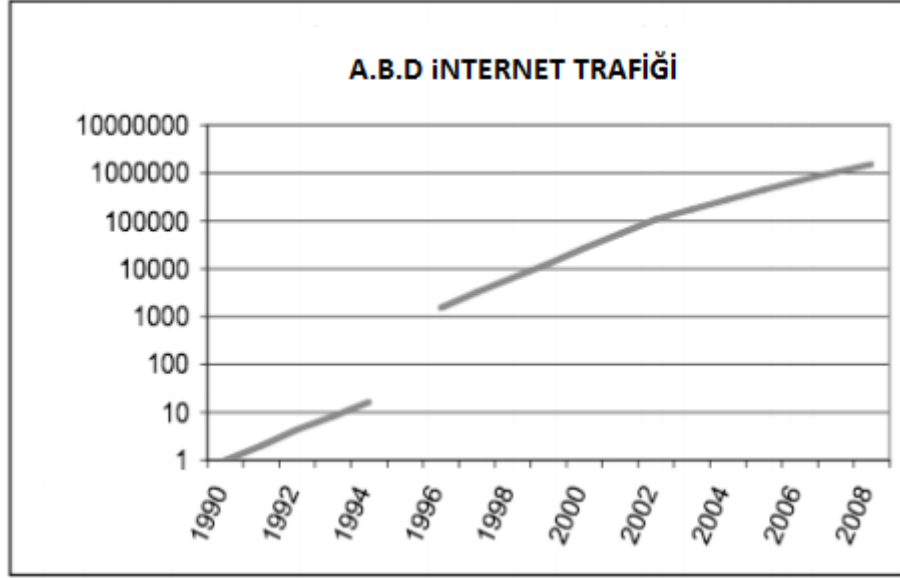
Web 2.0 internetin halk tabanına yayılması anlamında önemli işler başarmıştır. Dinamik bir yapıya kavuşan internet sitelerinin içeriğinin kullanıcılar tarafından oluşturulması, yapısal olarak internet yaşamını değiştirdi. İnternet siteleri artık sadece ortamı oluşturup, gerisini kullanıcıya bıraktı. Facebook, Flickr, Friendfeed, Myspace, Twitter, Youtube, Blogger, LinkedIn, Google+ gibi siteler bu temel üzerine kuruldu.



Şekil 2.1. Yeni nesil web

Sosyalleşen internet sınıfı burjuva internetini yıkarak artık ben de varım mesajını kuvvetli bir şekilde vermektedir. İnternet ortamında iletişim kuran kullanıcılar arasında oluşan veriler devasa boyuttadır. 2008 yılında yapılan bir araştırmaya göre internet ortamında bir ayda ortalama bir milyon terabyte veri taşınmakta ve bu sayı her geçen gün artmaktadır [16]. Bugün katılımcı, demokratik, paylaşımcı, üretken ve sosyal sıfatları ile

anılan Web 2.0 getirdiği doğrular yanında getirdiği zorluklar ile de baş etmek zorunda kalmıştır. Milyonların katılımı ile üretilen içerik devasa boyutlara ulaşmıştır.



Şekil 2.2. A.B.D Aylık Ortalama İnternet Trafik

Sosyal ağda yer alan aktörler ya da ilişkiler hakkında bir takım bilgileri çıkarsamak amacıyla yapılan, sosyal ağ yapısının incelenmesi işlemine sosyal ağ analizi adı verilmektedir. Sosyal ağ analizinde, ağdaki aktörlerin(düğümlerin) niteliklerinden çok aktörler arası ilişkilere odaklanılır. Örnek vermek gerekirse; bilimsel bir topluluktaki araştırmacıların performanslarını herhangi bir ölçüte göre (yayın sayısı gibi) önceden tahmin etmeye çalışacağımızı düşünelim. Geleneksel sosyal bilimler yaklaşımı, araştırmacıların yaşlarını, verdikleri burs miktarları, ait oldukları ekibin büyüklükleri gibi nitelikleri göz önünde bulundurur. Hemen ardından bu nitelikleri kullanarak yapılacak bir istatistiksel analiz ile istenilen ölçüt elde edilmeye çalışılır. Ancak ağ analizinde durum biraz farklı incelenir. Araştırmacıların özellikleri yerine topluluktaki bağlılıklarına odaklanılır. Örneğin, araştırmacıların sahip oldukları ilişkiler ve bu tür ilişkilerin araştırmacıların çalışmalarına sağlayacağı potansiyel fayda ve sınırlamalar üzerinde durulur. Belirli türde ilişkilerin düzenli olmasının performansı olumlu etkilediği yönünde sonuçlar çıkartılabilir ve hatta sadece bireysel performanslar değil, bunların ağa olan etkileri de değerlendirilebilir [17].

Sosyal ağ analizinde ağın yapısını anlamak için bağlantılardan yola çıkarak yapılan inceleme çalışması sonucunda, aktörlerin konumları ve bağlantı dereceleri, ağın büyüklük ve yoğunluğu, ağdaki gruplaşmalar hakkında bilgiler elde etmek mümkündür. Bu bilgiler, daha çok matematiksel yöntemlerin, ağı temsil eden veriler üzerinde çalıştırılması ile elde edilen sayısal verilerdir. Bu sayısal analiz verileri yorumlanıp, ilgili kullanım alanında yeni bilgiler elde edilmesinde değerlendirilebilir. Analiz sonuçlarının yorumlanması ağdan ağa farklılık gösterebilir. Bunun dışında ağ verisini görselleştirerek üzerinde görsel analiz çalışması yapılmasını sağlayan uygulamalar da mevcuttur.

Sosyal ağ analizi günümüzde pek çok alanda kullanılmakta olup bunların başında birey ve sosyal grup yapılarının ve davranışlarının incelenmesi (bileşenlerine ayırma, kümeleme, ilişkilerin belirlenmesi), elektronik ticaret ve çevrimiçi reklamcılık (müşteri profilinin çıkarılması ve eğilim analizi, kişiye özel reklamcılık ve teklif sunma), fiziki yapıların analizi (ulaşım, tesisat, altyapı) ve büyük veri kümelerinin analizi (medya takibi, akademik yayın analizi, genetik araştırmalar) gelmektedir.

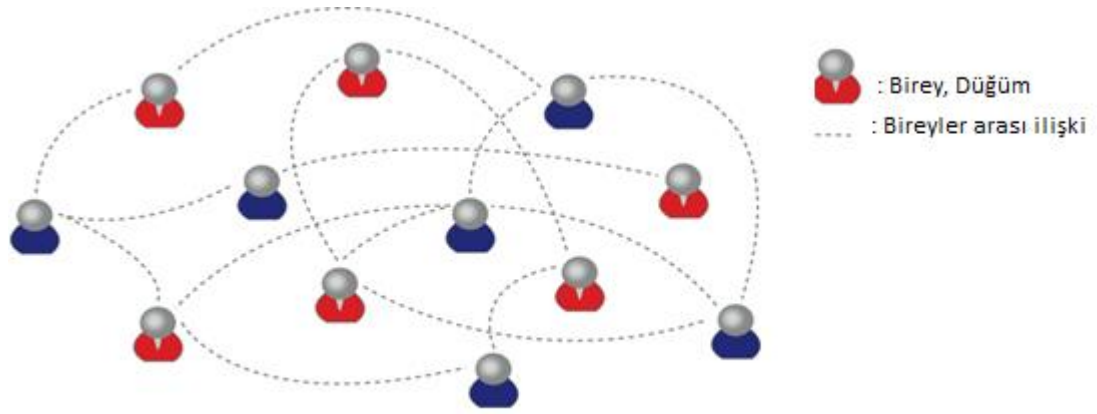
Sosyal ağ analizi yeni 19. yy. sonlarından itibaren sosyoloji biliminin öncülüğünde gelişim göstermiş bir yapıdır. Bilişim teknolojisinde yaşanan büyük gelişmelerin, bireylerin ve toplulukların etkileşimini çok farklı boyutlara taşıması ile daha da gelişmiştir.

Sosyal ağ analizinin günümüzde kullanım alanı hızla genişlemektedir. Başlangıçta çoğunlukla birey ve sosyal grup yapılarının ve davranışlarının incelenmesinde kullanılırken günümüzde ekonomi, ticaret, sağlık ve bankacılık gibi çok daha karmaşık alanlarda da kullanılmaktadır. 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 (SAA) yardımıyla nasıl ortaya çıkarıldığı bilinmektedir.

Sayısal ortamlarda iletişim ağından elde edilen somut 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 pratik 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

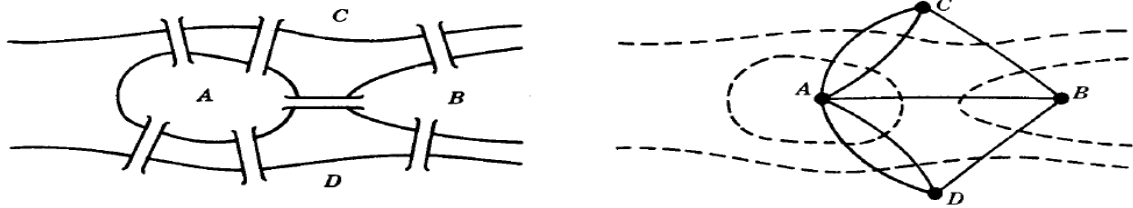
2.1. Çizge Teorisi

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

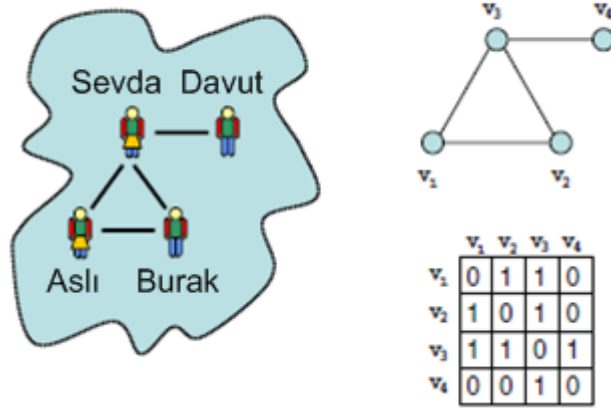
Çizge teorisinin çıkış noktası 18. yüzyıl Doğu Prusya kasabasıdır. Königsberg’in 2 adası ve 7 köprüsü vardır. Königsberg halkı ünlü matematikçi Euler’e, bir kişinin herhangi bir yerden başlayıp herhangi bir yerde durarak ve her köprüyü bir ve en fazla bir kez geçerek bir gezinti yapıp yapamayacağını 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 düğümlerinin komşu olduğu söylenir. Böylesi bir durumda U ve V , E ’nin uç noktaları olarak adlandırılır ve E ’nin U ve V ’yi bağladığı söylenir.

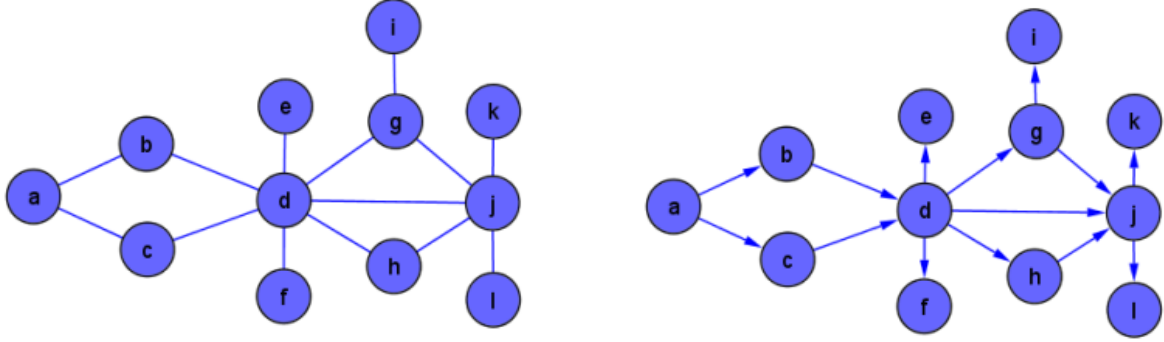
Şekil 2.5’de gösterildiği gibi V kümesindeki her v düğümü bir nokta (ya da küçük çember) ile temsil edilir ve her $E = \{v_1, v_2\}$ kenarı, v_1 ve v_2 uç noktalarını bağlayan bir çizgi ile gösterilir.



Şekil 2.5. Düğümler, kenarlar ve kenarların düğüm kümeleri

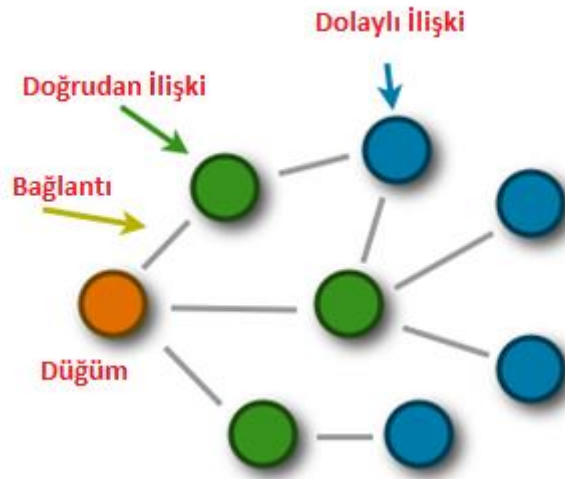
Çizge yapısının temsil edilmesi için genel olarak iki farklı yöntem tercih edilebilir. Bunlardan ilki, komşuluk listesi adı verilen, ağdaki her bir düğüm için komşu ya da bitişik olduğu düğümlerin listesinin dizi şeklinde gerçekleştirilmesidir. Bir düğümün komşuları liste dolaşarak tespit edilir. İkinci yöntem ise sık kullanılan komşuluk matrisi yöntemidir. Bu yöntemde çizgedeki n tane düğüm için oluşturulan $n \times n$ boyutundaki matris üzerinde birbirine komşu olmayan düğümlerin girdileri 0, komşu düğümlerin girdileri 1 olarak atanır. Matris işlenirken de her bir girdinin değerine bakılarak, karşılık geldiği satır ve sütundaki düğümler arasında kenar(bağlantı) olup olmadığına karar verilebilir.

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

Sosyal ağlar görselleştirilmiş birer veri haritaları gibidir. Aktörlerin birbirleriyle kurduğu ilişkilerden meydana gelmektedir. Bu yapı sayesinde oluşan veriler çizge kuramına göre yapılandırılabilir. Kişiler yani sosyal ağdaki üyeler birer düğüm olarak düşünülebilir. Şekil 2.7’ 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 2.7. Sosyal ağın çizge yapısı

3. BAĞLANTI TAHMİNİ

Günümüzde uygulama alanları giderek artış gösteren sosyal ağlar dinamik yapıya sahiptir. Bu yüzden üzerinde tahminlerde bulunmak zordur. Sosyal ağın merkezi denilebilecek düğümlerinin gelecekte de aynı merkeziliğe sahip olması beklenir. Fakat bu durumun tam tersi olabileceği de hesaba katılmalıdır. Bu ağlarda gelecekte yeni ilişkilerin oluşacağı veya var olan ilişkilerin yok olacağını tahmin etmek oldukça zor bir iştir. Sosyal ağ yapısıyla paralel olarak bu ağ üzerinde tahmini yapılacak niteliği de doğru seçmek gerekir. Bu da önemli bir sorundur. Yaşadığımız bilgi çağı toplumunda bu ağlardaki ilişkiler sayesinde bireylerin sosyal davranışlarını inceleyebilmekte, insan ilişkileri konusunda nitel ve nicel değerlendirmelerde bulunabilmekte ve bunlardan çok yararlı bilgiler elde edilebilmektedir [22].

3.1. Bağlantı Tahmini Problemi

Sosyal ağlarda bağlantı tahmini problemi, ağın gelecekteki yapısıyla ilgili önsezilerde bulunulmasıdır. Bağlantı tahmini yöntemleri sosyal ağdaki düğümlerin ve ilişkilerin özellikleri kullanılarak geliştirilmiştir. Ağdaki veriler düğümlerle, ilişkiler ise bağlantılarla tanımlanarak çizge yapısında ifade edilmektedir.

Sosyal ağlarda her düğüm vektörel yapıya sahip verilerdir. Veriler çeşitli ön işleme adımlarından geçtikten sonra tablolara dönüştürülür. Oluşturulan tablolarda satırlar değerleri, sütunlar ise düğümlerin özelliklerini belirtmektedir. Her bir düğümün niteliklerinin bilindiği ağlarda daha oluşmamış bağlantılar tahmin edilebilmektedir. Düğümler arası ilişkiler ve düğümlerin özellikleri kullanılarak düğümler arasındaki bağlantılar tahmin edilebilmektedir. Ağda ileride yeni ilişkiler oluşma ve ağa yeni düğümlerin katılma ihtimallerinin yanı sıra ağdan düğümlerin çıkma veya ilişkilerin eksilme ihtimalleri de bulunmaktadır.

Ağlar dinamik yapıya sahip olduğu için bu tahminlerin yapılması oldukça zordur. Var olan ağ yapısının resmi çekilmektedir. Yakın gelecekte bu ağa yeni bireylerin dâhil olup olmayacağının, yeni bağlantıların oluşup oluşmayacağı veya var olan bağlantıların ileriki zamanlarda da devam edip etmeyeceğinin hesaplanması gerekmektedir. Ağdaki bilgilerin nasıl tanımlanacağı da önemli bir sorundur. Var olan bilgilerin nasıl

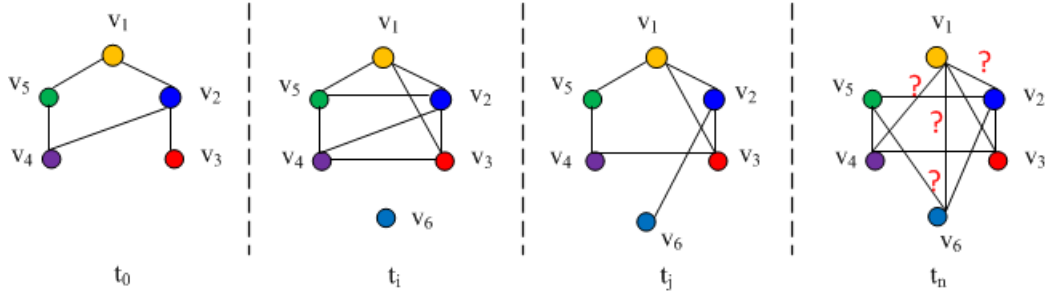
kullanılacağıının, doğru hesaplamalar (tahminler) yapmada ne kadar etkili olacağı iyi düşünölmelidir.

Bağlantı tahmininde ağın topolojik özelliklerinin yanında ağdaki düğümlere ve ilişki ait özellik türlerine de ihtiyaç duyulmaktadır. Karmaşık olmayan seyrek ağlarda bu bağlantı tahmini yapılırken bu bilgiler fayda sağlamaktadır. Bu şekilde daha verimli sonuçların alınacağı tahminler yapılmaktadır. Bazı durumlarda ise veri setleri dengesiz durumlar ortaya çıkabilmektedir. Düğüm sayıları ve bağlantı sayıları arasında tutarsızlıklar oluşabilmektedir. Bu durum dengeli verilerde kullanılan birçok modelin etkinliğini engellemektedir. Geniş sayıdaki düğümler ve/veya kenarlar çizgede ölçeklenirse, hesaplama yapan modellerin verimli olması şarttır. Yani oluşturulan modelin büyük sistemlerde kullanılmasının zorluğu ön plana çıkmaktadır [23].

Bağlantı tahmini yapılırken bağlantının oluşumu, türü ve ağırlığı keşfedilip buna göre tahminin doğruluğu analiz edilmelidir.

Bağlantı tahminiyle gelecekte sosyal ağda iki düğüm arasında bir ilişki kurulup kurulmayacağıının 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 [24].

Bağlantı tahmini probleminde, düğümler V (veri örnekleri) olarak düşünölürse $V = \{v_i\}_{i=1}^n$, 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 [25]. Şekil 3.2’de oluşan ağın zamansal değişimi yapılan ölçümlerle karşılaştırılarak tahminin doğruluğu ölçülebilmektedir.



Şekil 3.1. Sosyal ağın zamansal değişimi

3.2. Yapılan Bazı Çalışmalar

Bağlantı tahmini, sosyal ağların yaygınlaşmasıyla birlikte birçok alanda kullanılmaya başlanmıştır. Sosyal ağdaki bağlantılar kullanılarak varlıkların sınıfları ve özellikleri tahmin edilebilir. Ağdaki ilişkisel yapılar keşfedilip kümeleme ve analiz yapılabilir.

Bağlantı tahmini alanında yapılan birkaç çalışmaya bakacak olursak [26];

- Murata ve Moriyasu [27] İlan Tahtaları Soru Yanıtlama sosyal ağlarına Liben-Nowell ve Kleinberg modelini uyguladılar.
- Seglen [28] geniş etkisi olan faktörlerle dergilerdeki makalelerin eğilimlerini analiz etti.
- Shibata ve arkadaşları [26] farklı sosyal ağlar için farklı öğrenme yaklaşımları ve farklı özelliklerle çeşitli bağlantı tahmini modelleriyle atıf tahmini önerileri yaptılar.
- Vinkler ve Davidson [29] büyüyen dergilerdeki makalelerin atıf yapılma olasılığının daha yüksek olduğunu gösterdiler.
- Popescul ve Ungar [30] İstatistiki İlişkisel Öğrenme için bağlantı tahmini modelini inşa eden yeni bir yaklaşım önerdiler.
- Caragea ve arkadaşları [31] kümelenme yaklaşımına dayalı Live-Journal'da potansiyel arkadaşlıkları tahmin eden bir algoritma geliştirdiler. Live-Journal kullanıcı etkileşimlerine dayalı bir sosyal ağ günlüğü servsidir.
- Liben-Nowell ve Kleinberg [32] büyük ortak yazar ağları için bağlantı tahmini modeli önerdiler.
- Lu ve arkadaşları [33] iki düğüm arasında bir bağlantı varlığı olasılığını tahmin etmek için yerel bir yol endeksi sundular.

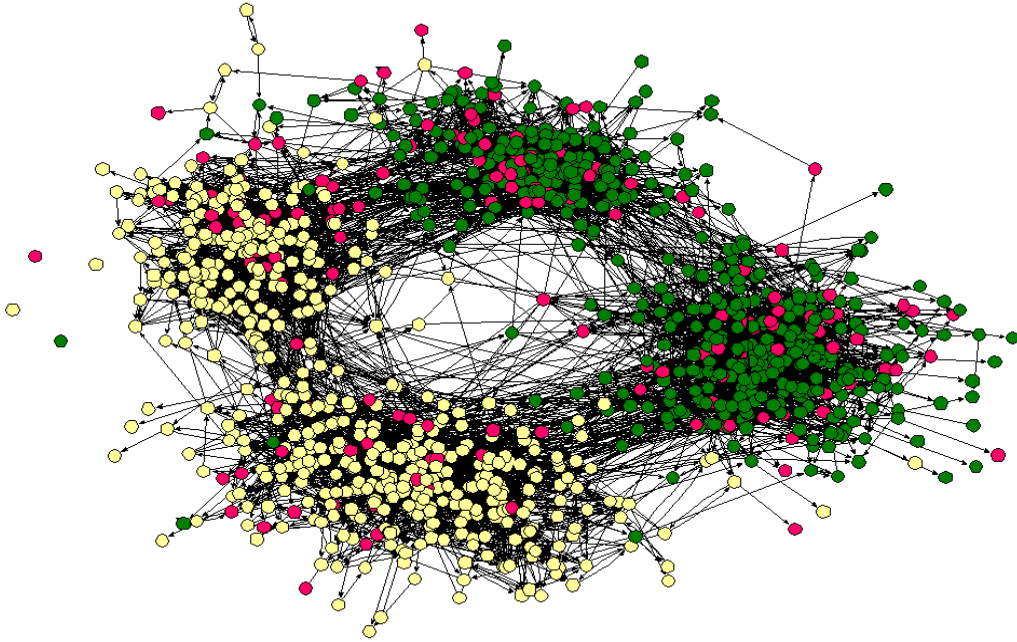
- Hasan ve arkadaşları [3, 34] birkaç denetimli öğrenme yöntemini bağlantı tahmini için test ettiler. Bu modeller karar ağaçları, k-en yakın komşu, çok katmanlı algılama, destek vektör makinesi, radyal tabanlı fonksiyon ağı ve bunlar arasındaki en başarılı sonuçlar Destek Vektör Makinesi modeli ile olmuştur.
- Clauset ve arkadaşları [35] sosyal ağların hiyerarşik yapısını araştırarak kısmen bilinen ağların eksik bağlantılarını yüksek doğruluk ile tahmin ettiler.
- Hwang ve arkadaşları [1] ortak yazarlı ağlar için tavsiye motorları önerdiler.
- Davis ve arkadaşları [1, 2] hastanın geçmiş bilgilerinden hastalık tahmini yapmak için yöntemler önerdiler.
- Folino ve arkadaşları [4] hastalık riskini tahmin eden bir yaklaşım sundular.
- Folino ve Pizzuti [5] hastalık ağlarında bağlantı tahmini yöntemlerini kullanan yaklaşımlar sundular.

3.3. İlişki ve Ağ Türleri

Sosyal ağların hayatımızdın birçok alanda yer almaktadır. Bu ağların popülerliği sayesinde bu ağlarda arkadaşlık, iş, evlilik, vb. birçok ilişki anlamlandırılabilir. Örneğin, Şekil 3.3'te ki etnik kökenlerine göre bir okuldaki arkadaşlık ağı birçok anlam barındırmaktadır [36]. Bu çeşitliliğe bakıldığında bunlar belirli ilişki kategorilerine ve teknik özellik profillerine göre sınıflandırılabilir. Burada kısmen Borgatti ve arkadaşlarının [37] çalışmalarından elde edilen sınıflandırmaları kullanmak uygulamaya göre seçilecek yöntemler konusunda yardımcı olacaktır. Bu sınıflandırmanın özeti Tablo 3.1' de verildi [38].

Tablo 3.1. Sosyal ağ arařtırmalarında karşılaşılan baē türleri ve veri seti özellikleri

Baē türü	Örnekler	Çizge/veri seti özellikleri
Benzerlikler	Aynı yerde yaşama Aynı kulüp/etkinliğe gitme Aynı cinsiyetten olma Benzer alışkanlıkları olma	İki mod ve yönlü kullanmadan önce dönüřtürme gerekir dönüřtürme sonucu yönsüz ilişki çoēu zaman aēırlıksız
İliřkiler	Akrabalık, evlilik, arkadaşlık, iş iliřkisi, Sevmeye/sevmeme	Çoēu zaman yönsüz çoēu zaman aēırlıksız
Etkileřimler	Görüşme, yardım etme, tavsiye verme	Genellikle yönlü ve aēırlıksız
Akışlar	Bilgi akışı, řirketler arası personel akışı, ülkeler arası ticaret	Genellikle yönlü ve aēırlıklı

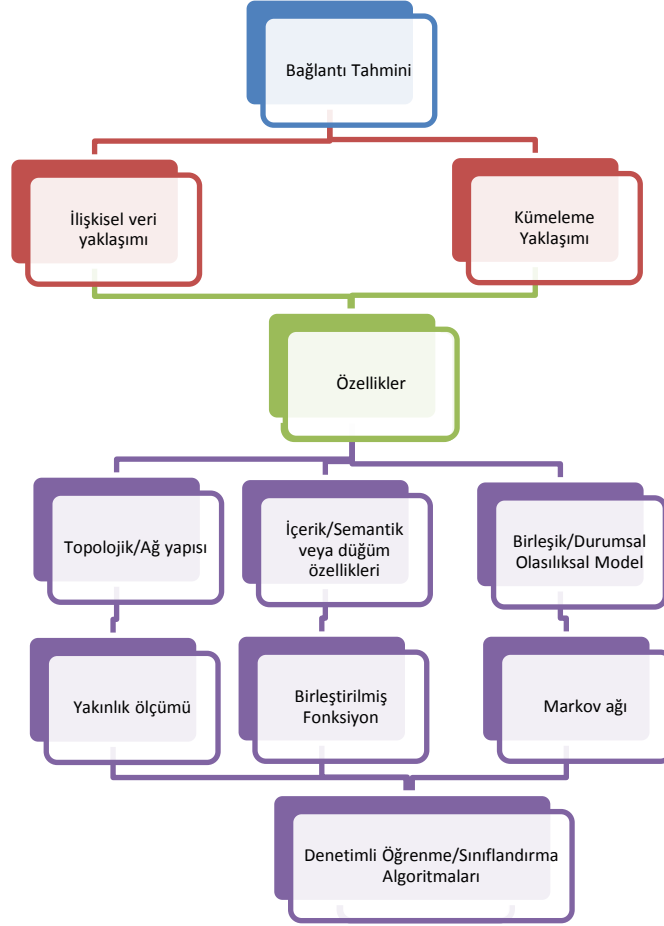


Şekil 3.2. Amerika'daki bir okuldaki arkadaşlık aēı. Renkler etnik kökenlerini göstermektedir.

3.4.Aēırlıklı ve Aēırlıksız Aēlarda Baēlantı Tahmini Yöntemleri

Sosyal aē analizi ve baēlantı tahmini yöntemleri birlikte kullanılarak sosyal aēlardan isteēe göre birçok bilgi elde edilebilmektedir. Bu iki yöntem verinin etkili bir şekilde bilgiye dönüşmesinde önemli fayda sağlamaktadır. Örneēin, borsa verileri sosyal aē analizi ve baēlantı tahmini yöntemlerinden geçirilerek borsanın geleceēiyle ilgili önemli bilgiler elde edilebilir. Aynı şekilde, market analizlerinde de etkili bilgiler elde edilip geleceēe

yönelik değişiklikler yapılabilir. Bağlantı tahmini ile birbirine bağlı birden çok ilişkisel tablodan yeni bilgilerde önerilebilmektedir.

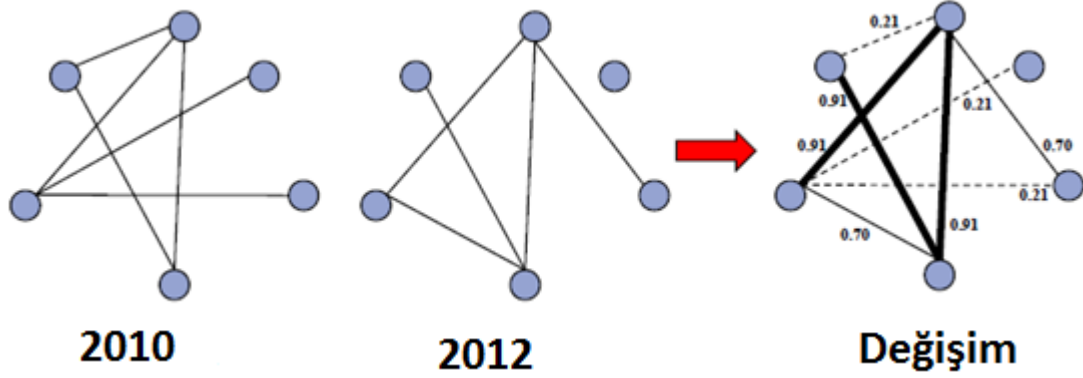


Şekil 3.3. Bağlantı tahminine farklı yaklaşımlar [39].

Bağlantı tahminini yapılırken kullanılan veri yapısı, oluşturulan ağ yapısı ve probleme göre farklı yaklaşımlar geliştirilebilir. Bu durum Şekil 3.4 ile görselleştirilmiştir.

Ağda iki düğüm (varlık) arasında ilişki kurulup kurulmadığını tanımlanmaya çalışıldığında bu düğümlerin ve aralarındaki ilişkinin özellikleri kullanılır. düğümlerin özellikleri, komşu sayısı, cinsiyet, etnik köken, yaş, başlık, coğrafi konum, vb. olabilmektedir. İlişkilerin özelliği ise, ilişki ağırlığı, yönü, vb. 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 3.5’de görüldüğü gibi farklı tarihteki ilişkilere göre yeni ilişkilerin oluşma ihtimali hesaplanabilmektedir.



Şekil 3.4. Ağ yapısındaki ilişkilerin zamansal değişimi [25].

Sosyal ağlarda ilişki tahmini için Denetimli ve Denetimsiz olarak iki yöntem kullanılır.

Denetimli yöntemler iyi tanımlanmış veya kesin bir hedef olduğunda kullanılır. Denetimli yöntemlerde, bir öğretmen tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir.

Denetimsiz yöntemler elde edilmesi istenilen sonuç için özel bir tanımlama yapılmamışsa veya belirsizlik söz konusu ise denetimsiz ifadesi kullanılır. Denetimsiz yöntemlerde, kümeleme analizinde olduğu gibi ilgili verilerin gözlenmesi ve bu verilerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır.

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 [25].

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.

3.5.Benzerlik Ölçütü Yöntemleri

Bu bölümde link tahmininde kullanılan ağırlıklı ya da ağırlıksız benzerlik ölçütlerinden bahsedilmektedir. Aşağıda ki ifadeleri anlamak için belirli açıklamalarda bulunmak gerekir.

- $\Gamma(x)$, sosyal ağdaki x düğümünün komşularının kümesidir.
- $|\Gamma(x)|$, sosyal ağdaki x düğümünün derecesi (komşularının sayısı) dir.
- $w(x, y)$, sosyal ağdaki x ve y düğümlerinin arasındaki ağırlıktır.

Yapılan hastalık tahmini çalışmamda yönsüz çizge teoremi kullanılmıştır ve benzerlik ölçütleri hesaplanırken düğümlerin kendileriyle bağlantılarının olmadığı kabul edilmiştir.

- Yönsüz çizge teoreminde $w(x, y) = w(y, x)$ olduğu kabul edilir.

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.

Ağırlıksız ağlar için Ortak Komşu Benzerlik Ölçütü denklem 3.1' deki gibi hesaplanır:

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

Ortak Komşu benzerlik ölçütü basitliğinden dolayı en yaygın kullanılan ölçütlerden biridir [3]. Bu ölçüte göre eğer iki düğüm arasında ne kadar çok ortak komşu varsa bunların ilerde birbirleri arasında bağlantı oluşması da o kadar olasıdır [40].

Ağırlıklı ağlar için Ortak Komşu Benzerlik Ölçütü denklem 3.2' deki gibi hesaplanır:

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \omega(x, z) + \omega(y, z) \quad (3.2)$$

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.

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

Ağırlıksız ağlarda Jaccard Benzerlik Ölçütü denklem 3.3'deki gibi hesaplanır :

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

Jaccard Benzerliği veri madenciliğinden çıkartılmış bir benzerlik ölçütüdür [mak13]. İki düğümün sahip olduğu ortak komşuları tüm komşularına oranlayarak daha yüksek oranlı bir değer elde edilir.

Ağırlıklı ağlarda Jaccard Benzerlik Ölçütü denklem 3.4' deki gibi hesaplanır:

$$S_{xy} = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{\omega(x, z) + \omega(y, z)}{\sum_{a \in \Gamma(x)} \omega(a, x) + \sum_{b \in \Gamma(y)} \omega(b, y)} \quad (3.4)$$

Tercihli Bağlantı İndeksleme: Ölçeksiz büyüyen ağlarda kullanılmaktadır. x düğümüne yeni bir bağlantının oluşturulma olasılığı $\Gamma(x)$ ile orantılıdır. Aynı yapı büyüme olmayan bir ağda belirli zaman dilimlerinde eski bir bağlantının kaldırılması ve yeni bir bağlantının oluşturulması olasılığı düğümün toplam komşu sayısı ile orantılı yapılır. x ve y düğümleri arasında bağlantı oluşma olasılığı $\Gamma(x) * \Gamma(y)$ ile orantılı olur. Bir düğümün bağlantı sayısı ne kadar yüksekse ilerde yeni bir bağlantı oluşturma olasılığı da o kadar yüksektir [makaleSuper]. Barabasi and Bonabeau [54], ve Newman [53] ‘e göre iki düğüm arasında bağlantı oluşma ihtimalinin düğümlerin ağdaki komşu sayılarıyla doğru orantılı olduğunu savunmuşlardır.

Ağırlıksız ağlarda Tercihli Bağlantı Benzerlik Ölçütü denklem 3.5’ deki gibi hesaplanır:

$$S_{xy} = \Gamma(x) * \Gamma(y) \quad (3.5)$$

Ağırlıklı ağlarda Tercihli Bağlantı Benzerlik Ölçütü denklem 3.6’ daki gibi hesaplanır:

$$S_{xy} = \sum_{a \in \Gamma(x)} \omega(a, x) * \sum_{b \in \Gamma(y)} \omega(b, y) \quad (3.6)$$

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

Ağırlıksız ağlarda Adamic-Adar Benzerlik Ölçütü denklem 3.7’ deki gibi hesaplanır:

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(|\Gamma(z)|)} \quad (3.7)$$

Adamic-Adar benzerlik ölçütü Jaccard benzerlik ölçütünün değiştirilmesiyle formüle edilmiştir [55]. Bu ölçüt az komşulara sahip ortak komşulukların önemini artırır. Bu

şekilde, özel (ya da güçlü) bir ortak komşuya sahip olmak düğüm çiftlerinin ilerde kendi aralarında bağlantı oluşturacağı ihtimalini arttırır.

Ağırlıklı ağlarda Adamic-Adar Benzerlik Ölçütü denklem 3.8’ deki gibi hesaplanır:

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\omega(x, z) + \omega(y, z)}{\log(1 + \sum_{c \in \Gamma(z)} \omega(z, c))} \quad (3.8)$$

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.

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

Liben-Nowell ve arkadaşları [36] ve Zhou ve arkadaşları [42] yerel benzerlik indislerini kullanarak gerçek ağlar üzerinde kıyaslamalar yaptılar. k_z ortak komşu sayısıdır.

Yapılan hastalık bağlantı tahmini uygulamasında *Ortak Komşu (CN)*, *Adamic Adar (AA)*, *Jaccard (J)* ve *Tercih Bağlantı İndeksleme (PA)* yöntemleri kullanıldı.

4. HASTALIK AĞININ OLUŞTURULMASI VE ÖNERİLEN YÖNTEM

4.1. Giriş

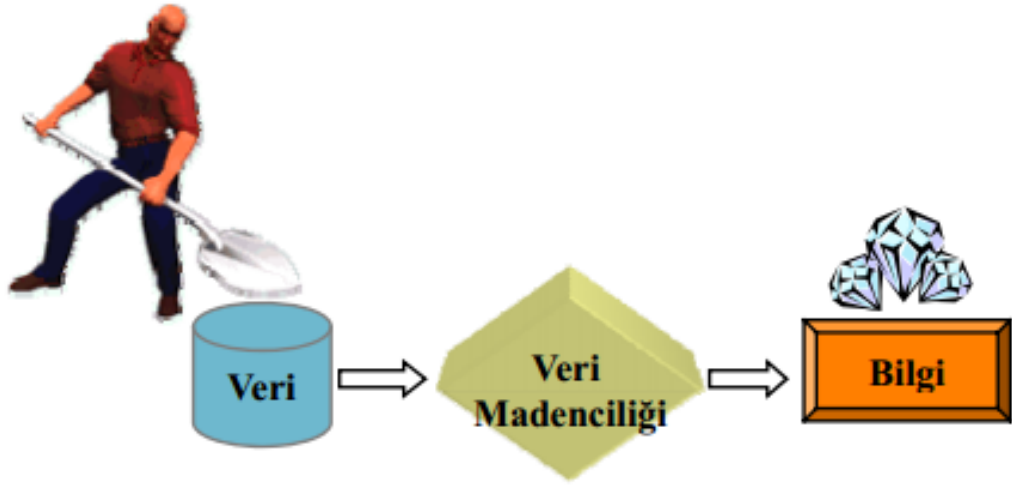
Bir önceki bölümde verilen yöntemlerin çalışmasını test etmek için bu bölümde bazı uygulamalar yapılmıştır. Bu kısımda uygulama tanıtılacak ve nasıl kullanılacağı anlatılacaktır.

Önerilen yöntem Java programlama dili kullanılarak geliştirilmiştir. Programın geliştirilmesi esnasında Netbeans IDE 7.4 ve JDK 7 kullanılmıştır. Yazılım veritabanı işlemleri için MySQL kullanılmıştır.

Geliştirilen uygulama dört aşamadan oluşmaktadır. Bunlar; Verinin Hazırlanması, Hastalık Ağının Oluşturulması, Komşuluk Düğümlerine Dayalı Bağlantı Tahmini Ölçütlerinin Kullanılması, Geliştirilen Sistemin Doğruluğunun Hesaplanması.

4.2. Verinin Hazırlanması

Çalışmada kullanılan veritabanı; 2013 yılında Fırat Üniversitesi Hastanesine gelip Hemogram Testi yaptıran hasta işlem kaydı bilgilerinden oluşmaktadır. Oldukça büyük olan ve dolayısıyla gereksiz bilgiler ihtiva eden bu veritabanı üzerinde çeşitli veri madenciliği veri hazırlama aşamaları kullanarak daha sade veriler elde edildi. Çünkü veri madenciliği, büyük hacimli veri yığınları içerisinde karar alabilmek için potansiyel olarak faydalı olabilecek, uygulanabilir ve anlamlı bilgilerin çıkarılmasına verilen addır. Veri madenciliği geniş anlamda veri analiz teknikleri bütünüdür ve tek başına bir çözüm değildir. Mevcut problemleri çözmek, kritik kararları almak veya geleceğe yönelik tahminleri yapmak için gerekli olan bilgileri elde etmeye yarayan bir araçtır.



Şekil 4.1. Veri Madenciliği

Verinin Hazırlanma Aşamaları:

- Başlangıçta “*Fırat Üniversitesi Hastanesi 2013 Yılı Hemogram Testi*” yaptıran 210.138 hasta işlem kaydı olan bir veritabanı bulunmaktaydı.

Tablo 4.1. Fırat Üniversitesi Hastanesi Hemogram Testi yaptıran Hasta Kayıt Verileri (2013)

	dosyaNo	adı	basvuruNo	altbirimAdı	tanılar
196606	785765		2713426	Nefroloji Kliniği	PROTEİN EKSİKLİĞİ ANEMİSİ D53.0
196607	785765		2713426	Nefroloji Kliniği	PROTEİN EKSİKLİĞİ ANEMİSİ D53.0
196608	785765		2713426	Nefroloji Kliniği	PROTEİN EKSİKLİĞİ ANEMİSİ D53.0
196609	785765		2713426	Nefroloji Kliniği	PROTEİN EKSİKLİĞİ ANEMİSİ D53.0
196610	824488		2713420	Göğüs Hastalıkları Polikliniği-2	ASTIM J45
196611	830380		2713414	Dermatoloji Polikliniği-2	AKNE L70,ANEMİ, DİĞER D64,TELOGEN EFFLUVİUM L65.0
196612	978795		2713403	Çocuk Nefroloji Polikliniği	ÜST SOLUNUM YOLU DİĞER HASTALIKLARI J39
196613	380717		2713398	Üroloji Polikliniği-1	ÜRİNER SİSTEMİN DİĞER BOZUKLUKLARI N39
196614	470222		2713375	Dermatoloji Polikliniği-1	KSEROSİS KÜTİS L85.3,TIRNAK BOZUKLUKLARI L60

- Tablo 4.1’ de bulunan veriler veritabanına aktarılırken Hasta *DosyaNO*, Hasta *Adı*, Hasta *BaşvuruNO*, Başvurduğu *AltbirimAdı*, Hastaya konulan *tanılar* sütunları kullanılmıştır.
- Bu sütunlardan oluşturulan veritabanında veri temizleme işlemleri yapılmıştır. Bu adımlar aşağıdaki gibidir.
 - i. Sadece bir kez hastaneye gelmiş olan hastalar temizlendi.
 - ii. Daha sonra birkaç kez hastaneye gelip aynı tanı konulan hastalar temizlendi.
 - iii. Bu temizleme aşamasından sonra birden fazla kez hastaneye gelmiş hasta kayıt sayısı: 168.006 oldu.
 - iv. Aynı hasta numarasına (hastaNO) sahip işlem kayıtları birleştirildi.

- Veri temizleme adımları bittikten sonra veritabanında 25.745 hasta dosyası ve bu hastalara konulan teşhisler kalmıştır.

4.3. Hasta Tanılarının ICD-10-Tr ile Kodlanması

ICD Kodları (*Hastalıkların Uluslararası Sınıflandırılması*) hastalıkların ve sağlık sorunlarının uluslararası platformda standart bir biçimde temsil edilebilmesini sağlayan istatistik kodlar bütünüdür. ICD-9 olarak bilinen önceki sürüm güncellenerek ICD-10 kodları oluşturulmuş ve günümüzde kullanılması sağlanmıştır. 1990 yılı Mayıs ayında 43. Dünya Sağlık Kongresi'nde onaylanan ICD-10 kodları 1994 yılında Dünya Sağlık Örgütü (WHO) üye ülkelerinde kullanılmaya başlamıştır. Kodlama sisteminin başlangıcı 1850'li yıllara kadar uzanmaktadır. Hazırlanan ilk liste Uluslararası Ölüm Sebepleri Listesi olarak 1893 yılında duyurulmuştur. Dünya Sağlık Örgütü, 1948 yılında bu listenin devam ettirilmesi sorumluluğunu üzerine almış ve 6. liste güncellemesiyle birlikte ICD Kodlarının temelleri atılmıştır.

ICD-10 kodları Türkiye Cumhuriyeti Sağlık Bakanlığı tarafından Türkçeleştirilerek kullanılmaya başlanmıştır.

Veri temizleme işlemleri tamamladıktan sonra veritabanında hastalara ait olan tanımlar **ICD-10-Tr** (*Hastalıkların Uluslararası Sınıflandırılması ve İlgili Sağlık Problemleri*), hastalıkların ve sağlık sorunlarının uluslararası sınıflama ölçütüne uygun olarak kodlanmıştır. Kodlama işleminden sonra veritabanı Tablo 4.2' deki gibidir.

Tablo 4.2. ICD-10-Tr sınıflama ölçütüyle kodlama aşamasından sonra oluşan veritabanı yapısı

dosyaNO	tanılar
1000009	E11.9,N39.0
1000017	G63.2,F41.9,G63.2,E78.4,N77.1,E11.9
1000025	N20.9,N21.0
1000029	K21,E78.4,I25,I21.9,R07.4
100004700	B18.1,B18.1,B18.1
1000049	N39.0,L50.9
100005868	M79.2,M75.4

Tablo 4.2’ de *dosyaNO* her bir hastayı, *tanılar* ise bu hastalara konulan hastalık teşhisleridir. Bu yapı elde edildikten sonra hastaların tanıları üzerinde yeni bir veri temizleme işlemi gerçekleştirildi. Örneğin; 100004700 dosya numarasına sahip hastaya üç kez de aynı teşhis konulmuştur. Bu hastanın oluşturulacak ağa bir katkısı olmayacağı için bu hastalar temizlendi. Diğer bir şekilde 1000017 hastasında da hastaya bir sonraki gelişinde tekrar G63.2 teşhisi konulmuştur. Böyle durumlarda ise, hastaya birden fazla kez aynı teşhis konulmuşsa bu teşhis sayısı bire indirilir.

4.4. Hastalık Ağının Oluşturulması

Bu çalışmada hastalıklar arası bağlantıları gösteren ağırlıklı bir hastalık ağı oluşturulmuştur. Bu ağı oluşturmak için;

- Hastalıklar arası ilişkileri tespit ederken her bir hastanın tanılarının kendi aralarında kombinasyonları elde edilir. Tablo 4.2’ deki verilerde son işlemler yapıldıktan sonra aşağıdaki yapı elde edilir:

Tablo 4.3. Hastalıklar arası ilişkiler

dosyaNO	tanılar	Hastalıklar Arası ilişkiler
1000025	N20.9,N21.0	(N20.9,N21.0)
1000086	F20,R07.4,R50.9,J18.9	(F20,R07.4);(F20,R50.9);(F20,J18.9);(R07.4,R50.9); (R07.4,J18.9);(R50.9,J18.9)
1000096	R42,R10.4,N39.0	(R42,R10.4);(R42,N39.0);(R10.4,N39.0)

- Tablo 4.3 ‘de oluşturulan yapı 2. Bölümde anlatılan çizge teorisine göre düzenlenmiştir. Bu teoriye göre düğümler arası ilişkiler $G(V,E)$ yapısı ile ifade edilir. Örneğin; (N20.9,N21) hastalık ilişkisinde N20.9 ve N21 düğümleri arasında bir ilişki olduğunu ifade eder.
- Çalışmada “*Yönsüz Ağırlıklı Hastalık Ağı*” oluşturulmuştur. Oluşturulan ağda 3.114 düğüm, ve bu düğümlere ait 66.223 bağlantı bulunmaktadır. Yönsüz olan ağda (a,b) ilişkisi (b,a) ilişkisine eşittir.

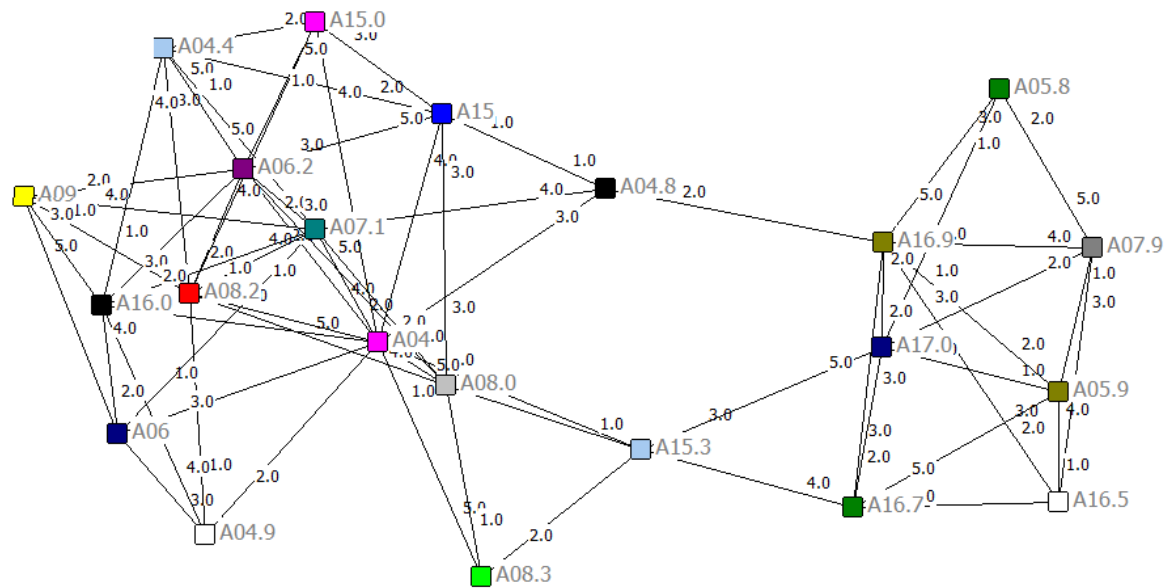
Sosyal ağların temsili için kullanılan veri formatı sosyomatrix (İng. sociomatrix) ya da yakınlık matrisi (İng. adjacency matrix) diye bilinen formattır. Bu matris yapısında hem satırlar hem de sütunlar incelendiğinde sosyal ağın bileşenlerine (hastalıklara) karşılık

gelistir. Çalışmadaki "Yönsüz Ağırlıklı Hastalık Ağı" verisi bu yöntemle aşağıdaki gibi temsil edilebilir:

Tablo 4.4. Ağırlıklı Hastalık Ağı ilk 9 hastalık sosyomatrisi

	A04	A04.4	A04.8	A04.9	A05.8	A05.9	A06	A06.2	A07.1
A04	0	0	0	0	0	0	0	4	2
A04.4	0	0	0	0	0	0	0	5	1
A04.8	3	0	0	0	0	0	0	0	4
A04.9	2	0	0	0	0	0	3	0	0
A05.8	0	0	0	0	0	0	0	0	0
A05.9	0	0	0	0	0	0	0	0	0
A06	3	0	0	0	3	0	0	0	5
A06.2	2	5	0	0	0	0	0	0	0
A07.1	5	0	0	0	0	0	1	3	0

Bu matristeki ilişkiler Şekil 4.2’ de görselleştirilmiştir. Bu matrisin yapısı söz konusu sosyal ilişkinin tüm özelliklerini yansıtmaktadır. Bu özelliklere bakılacak olursa:



Şekil 4.2. Yönsüz Ağırlıklı Hastalık Ağından ilk 22 Hastalık

- Matrisin diyagonal hücrelerinde (A04-A04, A08.0-A08.0, vb.) hep 0 değerini alır. Çünkü; söz konusu ilişki sosyal bir ilişkidir. Dolayısıyla hastalıkların kendi kendileriyle ilişkileri yoktur. Oysa ele alınan ilişki ekonomik, mesela "para transferi" ilişkisi olsaydı hesaplarım arası bir havale göndermem durumunda diyagonele sıfırdan farklı bir sayı yerleşmesi şaşırtıcı olmazdı. Benzer şekilde kimi

sosyal ilişkiler, örneğin "e-posta iletişimi", bireylerin kendilerine de ilişkisi olan bir durum ortaya çıkarabilir.

- Bu matris diyagonalinin çevresinde bir simetri gösterir. Yani A04.4-A16 hücresinde hangi sayı varsa, A16-A04.4 hücresinde de aynı sayı vardır. Çünkü; söz konusu ilişki *yönsüz* bir ilişkidir.
- Matrisin hücrelerinde 0 ve 1'den başka sayılarda vardır, çünkü bu ilişki *ağırlıklı* bir ilişkidir. Bir ilişkinin şiddeti/ağırlığı diğerinden farklı olabilir. Evlilik ilişkisi gibi ikili/ağırlıksız bir ilişki söz konusu olsaydı veri matrisimizde sadece 0 (ilişki yok) veya 1 (ilişki var) değerleri bulunacaktır.

4.5. Komşuluk Düğümlerine Dayalı Bağlantı Tahmini Ölçütleri

Bağlantı tahmini için geliştirilmiş birçok metot bulunmaktadır. Bunların çoğu benzerlik tabanlı algoritmalarlardır. Bazı ağlarda düğümlerin özellikleri bilinmediği için sadece ağ yapısı kullanılarak benzerlik katsayıları hesaplanır. Bu tez çalışmasında benzerlik tahmini yapılırken sadece ağ yapısı özellikleri kullanılmıştır. (x,y) düğümleri arasındaki bağlantı tahmini katsayısı $score(x,y)$ ile ifade edilir. X ve Y düğümlerinin aralarındaki bağlantı tahmini katsayısı ne kadar yüksekse ilerde bu iki düğüm arasında bağlantı oluşma ihtimali de o kadar yüksektir [5].

Çalışmada literatürde bulunan yerel benzerlik indekslerinden; *Ortak Komşu*, *Jaccard*, *Adamic-Adar*, ve *Tercihli Bağlantı* indeksleri kullanılarak bağlantı tahmini katsayıları hesaplanmıştır. Yerel benzerlik indeksleri düğümlerin komşuluk düğümlerine dayalı bağlantı tahmini ölçütleridir. Bu ölçütler 3. Bölümde ayrıntılı olarak anlatılmıştır.

4.6. Geliştirilen Yöntem

Oluşturulan Hastalık Ağı'ndan yapılan bağlantı tahminlerinin doğruluğunu hesaplamak için Folino ve Pizzuti'nin 2012 yılında yaptıkları [5] çalışmaya dayalı bir yöntem geliştirilmiştir.

- Öncelikle T veritabanı $T_{\text{eğitim}}$ ve T_{test} olarak iki parçaya ayrıldı.
- 3.114 düğümün ve 66.217 ilişkinin bulunduğu G ana ağdan, ağın merkezi düğümleri denilebilecek yüksek oranda komşu sayısına sahip olan düğümlerin

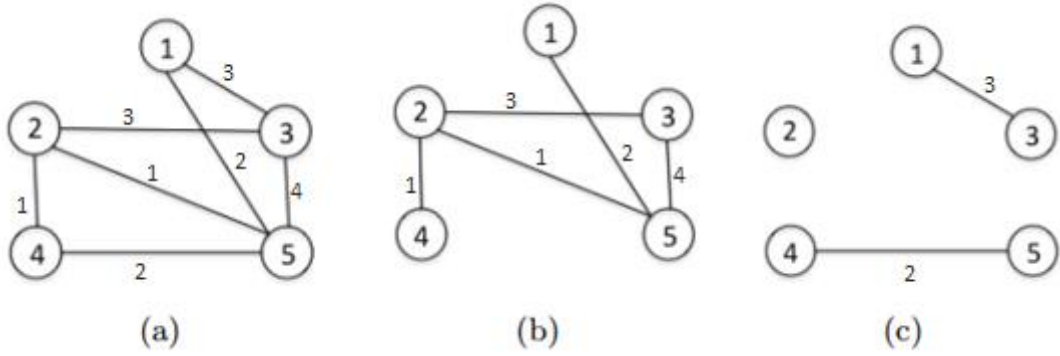
bağlantılarından rastgele 1000 tane ilişki çıkartılarak kalan 65.217 bağlantıyla $G_{eğitim}$ ağı, çıkartılan 1000 tane bağlantıyla da G_{yeni} ağı oluşturuldu.

- $E = E_{eğitim} \cup E_{yeni}$ ve $E_{eğitim} \cap E_{yeni} = \emptyset$ 'dir.
- Oluşturulan $G_{eğitim}$ ağı üzerinde eksik bağlantılar bulunup T_{test} veritabanına aktarıldı. $x, y \in (V \times V) - E_{eğitim}$ olan her bir düğüm çifti için bağlantı tahmini katsayısı hesaplandı. Önceki bölümde de bahsedildiği gibi x ve y düğümlerinin benzerlik katsayısı ne kadar yüksekse, gelecekte bu iki düğüm arasında bağlantı oluşma ihtimali de o kadar yüksektir.
- $G_{eğitim}$ ağında eksik olan tüm bağlantılar bulunup benzerlik katsayıları hesaplandıktan sonra benzerlik katsayısına göre azalan sıraya göre L_p listesine aktarıldı.
- Daha sonra E_{yeni} yani ağdan çıkartılan veri kümesinin sayısı n ise L_p listesinin top- n kadar yüksek katsayılı linkler seçilir. Daha sonra Bağlantı Tahmini doğruluğu aşağıdaki gibi hesaplanır;

$$Tahmin = \frac{L_p \cap E_{yeni}}{E_{yeni}} \quad (4.1)$$

Geliştirilen yöntem aşağıdaki örnekle uygulamalı olarak açıklayacak olursak;

Örnek: Şekil 4.3(a) beş düğüme, sekiz kenara ve üç tane var olmayan kenara ((1,2),(1,4),ve (3,4)) sahip bir G ağıdır. Bağlantı Tahmini Modelini uygulayabilmek için G ağında var olan bağlantılardan rastgele iki bağlantı seçilip $E_{yeni} = \{(1,3),(4,5)\}$ veri kümesi oluşturulur. G_{yeni} ağı Şekil 4.3(c)' de gösterilmiştir. Geriye kalan bağlantılar $E_{eğitim} = \{(1,5),(2,3),(2,4),(2,5),(3,5)\}$ veri kümesine aktarılıp Şekil 4.3(b)' deki $G_{eğitim}$ ağı oluşturulur.



Şekil 4.3. Bağlantı Tahmini Hesabında kullanılacak üç ağ: (a) G , (b) $G_{eğitim}$, ve (c) G_{yeni}

Yukarıdaki ağ yapıları elde edildikten sonra, $G_{eğitim}$ ağından 4.2 formülü kullanılarak eksik bağlantılar keşfedilir. N_l eksik bağlantılar, VxV ise Hastalık Ağı'nda bulunan tüm kenar ilişkileridir.

$$N_l = (VxV) - E_{eğitim} = \{(1,2),(1,3),(1,4),(3,4),(4,5)\} \quad (4.2)$$

Önceki bölümde bahsedilen *Jaccard*, *Ortak Komşu*, *Adamic-Adar*, *Tercih Bağlantılı* benzerlik ölçütleri kullanılarak Tahmin edilen bağlantıların katsayıları hesaplanır.

Son aşamadaki ağ ağırlıklı ağıdır.

Ağ yapısal olarak analiz edilip düğümlerin komşuları da her bir düğümün özelliği olarak düşünülerek bu benzerlik ölçütleri denklem 3.2, denklem 3.4, denklem 3.6 ve denklem 3.8 de verilen ağırlıklı ağlarda uygulanan ortak komşu, jaccard, tercihlili bağlantı ve adamic adar katsayısı kullanılarak hesaplanır.

Γ , $G_{eğitim}$ ağındaki düğümlerin komşuluklarını ihtiva eder. Her bir düğümün komşuları aşağıdaki gibidir:

$$\Gamma(1) = \{5\}, \Gamma(2) = \{3,4,5\}, \Gamma(3) = \{2,5\}, \Gamma(4) = \{2\}, \Gamma(5) = \{1,2,3\} \text{ 'dir.}$$

Tablo 4.5. Herbir eksik bağlantı düğümünün Jaccard skoru

Kenar (x, y)	skor(x, y)_J
(1,2)	0,43
(1,3)	0,55
(1,4)	0
(3,4)	0,33
(4,5)	0,25

E_{yeni} veri kümesinde ki eleman sayısı 2 olduğu için. Eksik bağlantı skorlarından en yüksek iki tanesi alınıp geliştirilen uygulamanın doğruluk oranı hesaplanır. Tablo 4.5’ te $G_{eğitim}$ ağından elde edilen olmayan ve ilerde olması tahmin edilen bağlantıların Jaccard benzerlik ölçütüne göre hesaplanan skorları bulunmaktadır. Bu tabloya göre en yüksek skorlu iki bağlantı $L_p^2 = \{(1,3), (1,2)\}$ ve $E_{yeni} = \{(1,3), (4,5)\}$ olduğuna göre geliştirilen yöntemin doğruluk oranı:

$$Tahmin = \frac{L_p \cap E_{yeni}}{E_{yeni}} = \%50 \text{ ‘dir.}$$

Yukarıda küçük bir ağ üzerine uygulanıp aşamaları detaylandırılarak anlatılan bağlantı tahmini yöntemi oluşturulan 3.114 düğüm ve 66.217 ilişkiye sahip ağa uygulandı. Bu uygulama sonucu elde edilen sonuçlar Bölüm 5’ te verilmiştir.

5. UYGULAMA SONUÇLARI

Bu bölümde oluşturulan hastalık ağı üzerinde uygulanan yöntemin sonuçları değerlendirilir.

5.1. Oluşturulan Ağın Yapısal Olarak İncelenmesi

Tablo 5.1. Yönsüz Ağırlıklı Hastalık Ağ 'ında ağın en merkezi düğümleri

Hastalık	Ağırlık
R10.4 (KARIN AGRISI)	23143
K21 (GASTRO-ÖZOFAJIAL REFLÜ HASTALIĞI)	21252
I10 (ESANSİYEL (PRİMER) HİPERTANSİYON)	17010
I25 (KRONİK İSKEMİK KALP HASTALIĞI)	11446
R07.4 (GOGUS AGRISI)	9859
I20.9 (GASTRO-ÖZOFAJIAL REFLÜ HASTALIĞI)	9818
N39.0 (ÜRİNER SİSTEM ENFEKSİYONU, YERİ TANIMLANMAMIŞ)	8944
M25.5 (EKLEM AGRISI)	6866
M79.1 (LATERAL EPİKONDİLİT)	6399
E78.4 (HİPERLİPİDEMİ-YÜKSEK KOLESTROL)	5708
R50.9 (ATEŞ)	5255
E11.9 (İNSÜLİN BAĞIMLI OLMAYAN DİYABETES)	4991
R06.0 (DİSPNE – NEFES DARLIĞI)	4463
I50 (KALP YETMEZLİĞİ)	4215
R11 (BULANTI VE KUSMA)	4156
K59.0 (KABIZLIK)	3992
D64 (ANEMİ)	3945
E55 (VİTAMİN D EKSİKLİĞİ)	3840
N30 (SİSTİT-MESANE ENFEKSİYONU)	3638
D50 (DEMİR EKSİKLİĞİ ANEMİSİ)	3629

Tablo 5.1, 3.114 düğümlü *Yönsüz Ağırlıklı Hastalık Ağı*'nın en merkezi düğümleri gösterilmiştir. Bu tabloya göre ağda en çok komşuya sahip, ağa en hakim düğüm R10.4 (Karın Ağrısı) hastalık düğümüdür.

Tablo 5.2. Yönsüz Ağırlıklı Hastalık Ağ 'ında ağırlık derecesi en yüksek 10 bağlantı

Sıra	Hastalık-1	Hastalık-2	Skor
1	I10	I25	1333
2	I20.9	I25	1290
3	K21	R10.4	1211
4	I10	I10.9	1076
5	I10	K21	893
6	I25	R07.4	776
7	N39.0	R10.4	771
8	E78.4	I25	741
9	I10	R07.4	651
10	I10	R10.4	628

Tablo 5.2, 66.217 ilişkiye sahip *Yönsüz Ağırlıklı Hastalık Ağ* 'ında birlikte görülme durumu en fazla olan diğer bir deyişle ağın en yoğun ilişkileri bulunmaktadır. Bu tabloya göre ağın en yüksek ağırlıklı ilişkisi Hipertansiyon ve Kronik İskemik Hastalığı arasındaki ilişkidir.

5.2. Bağlantı Tahmini Yönteminin Değerlendirilmesi

Yönsüz Ağırlıklı Hastalık Ağı geliştirilen bağlantı tahmini yöntemi kullanarak $G_{\text{eğitim}}$ ve G_{test} olmak üzere iki ağa ayrıldı. Oluşturulan $G_{\text{eğitim}}$ ağında olmayan fakat ilerde oluşması muhtemel 9.737.373 bağlantı keşfedildi. Çalışmada yönsüz ağ yapısı kullanıldığı için ilerde oluşacak bağlantılar yönsüz hale getirilip 3.654.212 bağlantıya düşürüldü ve L_p listesine atıldı. Keşfedilen bu bağlantıların; *Jaccard*, *Ortak Komşu*, *Tercih Bağlantılı* ve *Adamic Adar* benzerlik ölçütü algoritmaları kullanılarak benzerlik tahmini skorları hesaplandı. Daha sonra doğruluk oranı hesabıyla bu algoritmaların performansları değerlendirildi.

Tablo 5.3. Jaccard Bağlantı Tahmini Ölçütü Top-10 Listesi

Hastalık -1	Hastalık -2	Skor
K21	R10.4	0,9455456137657166
I10	I25	0,9431555271148682
I20.9	I25	0,9189950823783875
I10	N30	0,9137538
I20.9	K21	0,91070175
I25	J45	0,9095146
E13	E78.4	0,88658863
E11.9	N30	0,8583254
E78.4	M54.5	0,8549672
D50	R50.9	0,8539672

Tablo 5.4. Ortak Komşu Bağlantı Tahmini Ölçütü Top-10 Listesi

Hastalık -1	Hastalık -2	Skor
J45.9	E83.5	79197
K21	R10.4	38687
I20.9	K21	27189
K05	I20.9	26898
F82	J18.9	22865
I10	I25	22399
H81.1	S72	22364
F92.0	M81	21409
E55.9	R10.4	20917
I10	N30	18456

Tablo 5.5. Tercih Bağlantılı Bağlantı Tahmini Ölçütü Top-10 Listesi

Hastalık-1	Hastalık-2	Skor
I10	I25	2702579
K21	R10.4	2352648
I20.9	K21	2277442
I20.9	I25	1115970
E55.9	R10.4	1025219
I10	N30	950003
I25	R06.0	696788
E55	E78.4	548100
I10	K58	535914

Tablo 5.6. Adamic-Adar Bağlantı Tahmini Ölçütü Top-10 Listesi

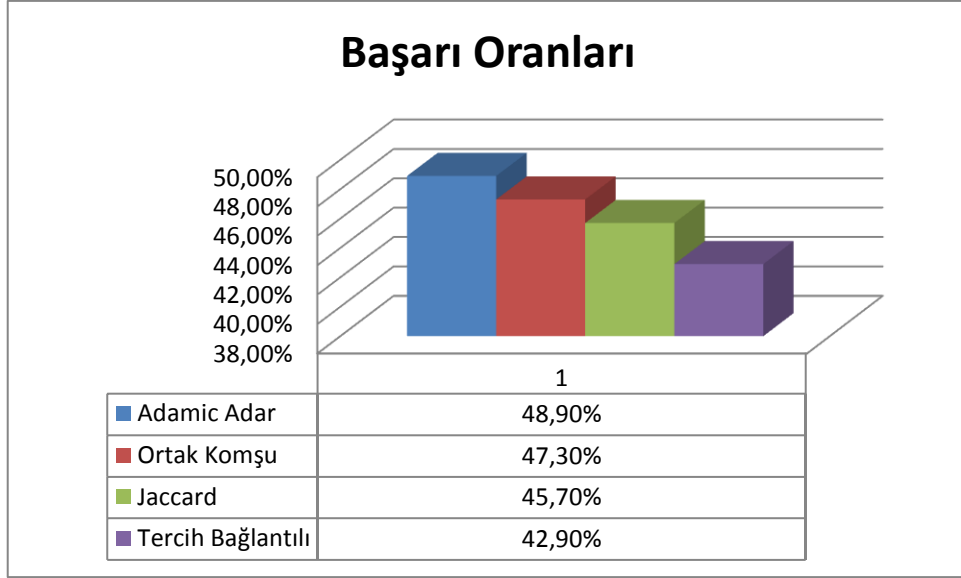
Hastalık - 1	Hastalık-2	Skor
K21	R10.4	3643,0918935666386
I20.9	K21	2639,9900826860508
I10	I25	2223,1512653563627
E55.9	R10.4	2059,1074592469104
I10	N30	1862,825665662026
I20.9	I25	1546,1524281578386
I25	R06.0	1490,9336473353599
H10.2	R10.4	1471,729871530179
I10	K58	1435,4668043274155
N30.9	R10.4	1377,762897878183

İlerde oluşması muhtemel olan bağlantı tahminleri için her bir benzerlik ölçütü algoritmasına göre skorlar hesaplanır. G_{yeni} ağında 1000 tane ilişki bulunduğu için L_p bağlantı tahmini listesindeki en yüksek skorlu 1000 bağlantı seçilip L_p^{1000} listesine atılır. Bu listeye göre bağlantı tahmini formülü uygulanarak geliştirilen uygulamanın başarı oranlarını hesaplanır.

.

Tablo 5.7. Geliştirilen Yöntemin Başarı Oranları

BAŞARI ORANLARI		
Benzerlik Ölçütü	Doğru Sayısı	Başarı Yüzdesi
Adamic Adar	489	%48,9
Ortak Komşu	473	%47,3
Jaccard	457	%45,7
Tercihli Bağlantılı	429	%42,9



Şekil 5.1. Başarı Oranları Grafiği

Hastaların kendi hastalık geçmişleri ve benzer hastaların hastalık geçmişlerini kullanarak oluşturulan Yönsüz Ağırlıklı Hastalık Ağı' nda bağlantı tahmini yöntemi kullanılarak analiz edilir. Elde edilen sonuçlara göre en yüksek başarı performansı %48,9 ile *Adamic-Adar* Benzerlik Ölçütüne aittir. İkinci sırada %47,3 ile *Ortak Komşu*, üçüncü sırada %45,7 ile *Jaccard* ve dördüncü sırada %42,9 ile *Tercih Bağlantılı* benzerlik ölçütü yer almıştır.

6. SONUÇ

Son yıllarda oldukça ilgi gören sosyal ağ analizinin kullanım alanları hızlı bir şekilde artmaktadır. Canlı ve cansız varlıklar arasında birçok yönden ilişki bulunmaktadır. Bu ilişkiler sosyal ağ analizi ile bilgi ağı şeklinde ele alınıp bu yapıların daha iyi anlaşılması sağlanır. Ağdaki aktörler arasında olası bağlantı tahmini, kullanıcıların davranışlarından örüntü çıkarımı, topluluk keşfi sosyal ağ analizi alanındaki çalışma konularından bazılarıdır.

Bağlantı tahmini konusu makine öğrenmesi alanında son yıllarda üzerinde çalışmaların artarak devam ettiği bir alandır. Günümüzde bibliyografik alan [43], moleküler biyoloji [44], adli soruşturmalar [10], 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.

Çalışmada sosyal ağlarda bağlantı tahmini problemi ele alındı ve sosyal ağlarda bağlantı tahmini yapacak bir yöntem önerildi. Yöntemde bağlantı tahmininde kullanılan benzerlik ölçütlerinden: *Jaccard*, *Adamic-Adar*, *Ortak Komşu*, *Tercih Bağlantılı* algoritmalar kullanılmıştır. Ağdaki düğümlerin aralarında ortak komşuları fazla olan düğümlerle bağlantı kurma ihtimali yüksektir. Fakat çalışma ağırlıklı ağ olduğu için bağlantı tahmini yapılırken sadece ortak komşuluklar değil aynı zamanda düğümler arasında var olan ilişkilerin ağırlıkları da hesaba katıldı. Önerilen yöntem benzer çalışmalarla kıyaslandığında iyi sonuçlar vermiştir.

Önerilen yöntem, Fırat Üniversitesi Hastanesinin biyokimya laboratuvarından alınan 2013 yılında kan testi yaptıran hasta verileriyle oluşturuldu. Ağ oluşturulurken bu verilerin bir kısmı ağı eğitmek için bir kısmı da analiz etmek için kullanılır.

Geliştirilen yöntemi uygulandıktan sonra oldukça verimli sonuçlar elde edilmiştir. Literatürde bilinen *Jaccard*, *Adamic-Adar*, *Tercih Bağlantılı* ve *Ortak Komşuluk* düğümler arasındaki yakınlık parametreleri kullanılarak daha önce var olmayan bağlantılar tahmin edilmiştir. *Adamic –Adar* benzerlik ölçütünde en yüksek 1000 tane bağlantı tahmininden 489 doğru tahmin edilmiş ve bu durum %48,9 ile en yüksek performanslı ölçüt olmasını

sağlamıştır. Adamic-Adar'dan sonra ikinci sırada %47,3 ile Ortak Komşu, üçüncü sırada %45,7 ile Jaccard ve dördüncü sırada %42,9 ile Tercih Bağlantılı benzerlik ölçütü yer almıştır.

KAYNAKÇA

- [1] Darcy. A. D. 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.
- [2] Darcy 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.
- [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] Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P. and Bhattacharjee, B., 2007, Measurement and analysis of online social networks, Proceedings of the 7th ACM SIGCOMM Conference on Internet measurement, San Diego, California, USA, 29-42 pp.
- [7] <http://www.bretswanson.com/index.php/2009/03/internet-update/>, Erişim Tarihi 02.06.2011.
- [8] Ögüt, S., 2009. "Veri Madenciliği Kavramı ve Gelişim Süreci", Görsel İletişim Tasarımı Bölümü, İletişim Fakültesi, Yeditepe Üniversitesi, İstanbul.
- [8] Milgram, S., 1967. "The Small-World Problem", Psychology Today, Vol. 1, No. 1,s. 61-67.
- [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 Alikılıç 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] 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, 234p.
- [18] Karıcı, 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] Volkova, S., 2009. "Link Prediction in Social Networks", KDD Lab Course Project.
- [23] Menon, A.K. ve Elkan, C., 2011. "Link Prediction via Matrix Factorization", ECML/PKDD 2, 437-452.
- [24] Getoor, L. ve Diehl C.P., 2005. "Link Mining: A Survey", SIGKDD Explorations, 7(2), 3-12.
- [25] Scott, J., 2000. "Social Network Analysis: A Handbook," Sage Publications, 209s.
- [26] 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.
- [27] Murata, T. ve Moriyasu, S., 2008. "Link Prediction based on Structural Properties of Online Social Networks", New Generation Computing, 26(3), 245–257.
- [28] 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.
- [29] Vinkler, P. ve Davidson, G.S., 2002. "Dynamic changes in the chance for citedness", Scientometrics, 54(3), 421–434.

- [30] 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.
- [31] 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).
- [32] 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.
- [33] 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.
- [34] Hasan, M.A., ve Zaki, M.J., 2011. "A Survey of Link Prediction in Social Networks", *Social Network Data Analytics*, Springer, s. 243-275.
- [35] Clauset. A, Moore, C. ve Newman, M.E.J., 2008. "Hierarchical structure and the prediction of missing links in Networks", Nature, 453, 98–101.
- [36] Moody J., 2001. "Race, School Integration, and Friendship Segregation in America", American Journal of Sociology 107, 679–716 .
- [37] Borgatti, Stephen P., Mehra A., Brass D.J. ve Labianca G., 2009. "Network Analysis in the Social Sciences", Science 323 (5916), 892–895.
- [38] Gençer, M., 2013. "Sosyal Ağlar", İstanbul Bilgi Üniversitesi.
- [39] Vartak, S., 2008. "A survey on Link Prediction", State University of New York.
- [40] Newman, M.E.J.: Clustering and preferential attachment in growing Networks. Phys.Rev. E 64, 025102 (2001)
- [41] Adamic, L.A., Adar, E.: Friends and neighbors on the web. Social Networks 25(3)(2003) 211-230
- [42] Barabasi, A. L., Bonabeau, E.: Scale-Free Networks. Scientific American 288(5)(2003) 60-69
- [43] 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.
- [44] Barabasi, A. L. ve Albert, R., 1999. "Emergence of scaling in random networks", Science 286, 509–512.

ÖZGEÇMİŞ

SERPİL GÜL

Elazığ – Türkiye

serpilgul90@gmail.com

KİŞİSEL BİLGİLER:

Doğum tarihi : 28.08.1990

Doğum Yeri : ELAZIĞ

EĞİTİM:

2008–2012

Fırat Üniversitesi Mühendislik Fakültesi

Bilgisayar Mühendisliği, Elazığ

2004–2008

Mehmet Akif Ersoy Yabancı Dil Ağırlıklı Lise, Elazığ

İŞ TECRUBELERİ:

Kurum adı:

Tunceli Üniversitesi Bilgisayar Mühendisliği Araştırma
Görevlisi (2014-Halen çalışıyorum)

Şirket adı:

Sonra Bilgi Teknolojileri (Elazığ, www.sonra.com.tr, 2012-
2013)

Sia Teknoloji Ltd. Şti (Elazığ, www.siateknoloji.com.tr,
2013-2014)