



**BÜYÜK ÖLÇEKLİ İKİ PARÇALI KARMAŞIK AĞLARDA
GÜÇLENDİRİLMİŞ YANSIMA MODELİNE DAYALI
BAĞLANTI TAHMİNİ YÖNTEMLERİ**

Serpil ASLAN

**Doktora Tezi
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Prof. Dr. Mehmet KAYA
EYLÜL-2018**

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

**BÜYÜK ÖLÇEKLİ İKİ PARÇALI KARMAŞIK AĞLARDA GÜÇLENDİRİLMİŞ
YANSIMA MODELİNE DAYALI BAĞLANTI TAHMİNİ YÖNTEMLERİ**

DOKTORA TEZİ

Serpil ASLAN

(141129202)

Anabilim Dalı: Bilgisayar Mühendisliği

Programı: Kuramsal Temeller

Danışman: Prof. Dr. Mehmet KAYA

Tezin Enstitüye Verildiği Tarih: 18 Eylül 2018

EYLÜL-2018

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**BÜYÜK ÖLÇEKLİ İKİ PARÇALI KARMAŞIK AĞLARDA GÜÇLENDİRİLMİŞ
YANSIMA MODELİNE DAYALI BAĞLANTI TAHMİNİ YÖNTEMLERİ**

DOKTORA TEZİ

Serpil ASLAN

141129202

Tezin Enstitüye Verildiği Tarih: 18 Eylül 2018

Tezin Savunulduğu Tarih: 12 Ekim 2018

Tez Danışmanı : Prof. Dr. Mehmet KAYA (F.Ü)

Diğer Jüri Üyeleri: Prof. Dr. İbrahim TÜRKOĞLU (F.Ü)

Doç. Dr. Taner TUNCER (F.Ü)

Prof. Dr. Ali KARCI (İ.Ü)

Dr. Öğr. Üyesi U. Baran BALOĞLU (M.Ü)

EYLÜL-2018

ÖNSÖZ

Sunulan doktora tez çalışmasında, gerçek dünya verilerinden elde edilmiş büyük ölçekli iki parçalı akademik bilgi ağlarında uygulanmak amacıyla benzerlik tabanlı yeni bir bağlantı tahmini yöntemi önerilmiştir. Önerilen yöntemde ağın evrimsel gelişiminin göz ardı edildiği ya da dikkate alındığı iki yeni yakınlık ölçütü algoritması önerilmiştir. Deneysel sonuçlar literatürde bilinen klasik yöntemler kullanılarak test edilmiştir. Uzun ve yorucu bir çalışmanın ürünü olarak ortaya çıkan bu çalışmada birçok kişinin katkısı bulunmaktadır.

Doktora çalışmalarım kapsamında bana kendisiyle çalışma fırsatı vererek önümde yeni ufuklar açan, manevi desteğini, bilgi ve tecrübelerini eksik etmeyen, akademik çalışma ve disiplini kazanmamı sağlayan, sonsuz sabrı ve anlayışı dolayısıyla değerli danışman hocam Prof. Dr. Mehmet Kaya'ya,

Tez çalışmamda M.F.17.01. nolu proje desteğinden dolayı FÜBAP'a,

Verdiği sonsuz sevgi ve moralle her zaman yanımda olan kıymetli eşime,

Tüm hayatım boyunca beni her konuda destekleyen, sevgi ve yardımlarını esirgemeyen aileme,

Sonsuz teşekkürlerimi sunarım.

**Serpil ASLAN
ELAZIĞ-2018**

İÇİNDEKİLER

Sayfa No

ÖNSÖZ	II
İÇİNDEKİLER.....	III
ÖZET	V
SUMMARY	VI
ŞEKİLLER LİSTESİ	VII
TABLolar LİSTESİ	IX
KISALTMALAR.....	X
SEMBOLLER LİSTESİ	XI
1. GİRİŞ	1
1.1. Tezin Amacı ve Katkıları.....	3
1.2. Tezin Yapısı.....	4
2. KARMAŞIK AĞLAR	5
2.1. Karmaşık Ağ Analizi	7
2.2. Çizge Teorisi.....	8
2.2.1. Matematiksel Tanımlar	12
2.3. Karmaşık Ağların Özellikleri	13
2.4. İki Parçalı Ağlar.....	14
2.4.1. İki Parçalı Ağların Tek Parçalı Ağlara Yansıtılması	16
2.4.2. İki Parçalı Ağların Ağırlıklı Tek Parçalı Ağlara Yansıtılması	17
3. BAĞLANTI TAHMİNİ.....	19
3.1. Bağlantı Tahmini Problemi.....	20
3.2. Bağlantı Tahmini Teknikleri.....	21
3.2.1. Düğüm Tabanlı Ölçütler	24
3.2.2. Topoloji Tabanlı Ölçütler	25
3.2.2.1. Komşuluk Tabanlı Ölçütler	26
3.2.2.2. Yol Tabanlı Ölçütler.....	29
3.2.2.3. Rastgele Yürüyüş Tabanlı Ölçütler	30
3.2.3. Sosyal Teori Tabanlı Ölçütler.....	32
3.2.4. Öğrenme Tabanlı Ölçütler	32
3.3. İki Parçalı Ağlarda Bağlantı Tahmini.....	33

3.3.1. İki Parçalı Ağlarda Bağlantı Tahmini Yöntemleri.....	35
3.3.1.1. İç Bağlantılar Tahmini Yöntemi.....	35
3.3.1.2. İşbirlikçi Filtreleme Yöntemi	36
3.3.2. İki Parçalı Ağlarda Bağlantı Tahmini Problemini Ele Alan Çalışmalar	37
3.4. Bağlantı Tahmini Değerlendirme Ölçütleri	39
4. İKİ PARÇALI AKADEMİK BİLGİ AĞI İÇİN GELİŞTİRİLEN BAĞLANTI TAHMİNİ YÖNTEMİ	43
4.1. Veri Toplama ve Temel Bilgi Çıkarımı	44
4.2. Büyük Ölçekli İki Parçalı Akademik Bilgi Ağının Oluşturulması	47
4.3. Yazar-Konu Akademik Bilgi Ağında Konu Tavsiyesi için Önerilen Bağlantı Tahmini Yöntemi.....	48
4.4. Önerilen Yöntemin Sözde Kodu.....	55
4.5. Uygulama Sonuçları	57
5. İKİ PARÇALI AKADEMİK BİLGİ AĞI İÇİN GELİŞTİRİLEN ZAMAN FARKINDALI BAĞLANTI TAHMİNİ YÖNTEMİ	65
5.1. Akademik Bilgi Ağının Oluşturulması	66
5.2. Zaman Farkındalı İki Parçalı Ağ Yapısı	66
5.2.1. Zaman Farkındalı İki Parçalı Akademik Bilgi Ağının Gelişen Yapısı	68
5.3. İki Parçalı Ağlarda Zaman Farkındalı Bağlantı Tahmini Yöntemi	69
5.3.1. Zaman Farkındalı Yakınlık Ölçütü Algoritması	69
5.3.2. İki Parçalı Ağlarda Zamansal Olaylar	70
5.3.3. İki Parçalı Ağlarda Zaman Farkındalı Yakınlık Ölçütü Algoritması	72
5.4. Önerilen Yöntemin Sözde Kodu.....	75
5.5. Uygulama Sonuçları	76
5.6. Farklı Çerçeve Boyutuna Sahip Ağlarda Durum Değişim Oranının Sonuçlara Etkisi	83
6. SONUÇLAR VE ÖNERİLER	88
KAYNAKLAR	91
ÖZGEÇMİŞ	100

ÖZET

Gerçek dünya verilerinden oluşturulmuş birçok karmaşık ağ, doğası gereği iki parçalı bir yapıya sahiptir. İki parçalı ağlar, gerçek dünya sistemlerinin karmaşık yapısı hakkında önemli fikirler veren, farklı düğüm grupları arasındaki etkileşimleri temsil eden karmaşık bir ağ türüdür. Son zamanlarda, özellikle büyük ölçekli, karmaşık ağlarda bağlantı tahmini çalışmaları farklı bilim alanlarındaki araştırmacılar için ilgi odağı haline gelmiştir. Birçok gerçek dünya sistemi gibi akademik bilgi ağları da iki parçalı bir yapıya sahiptir. Akademik bilgi ağları sürekli değişen ve gelişen bir yapıdadır. Bu ağlarda, yazarlara kendi alanlarına uygun popüler konu başlıklarını tavsiye eden bir sistemin bulunması araştırmacıların işini kolaylaştıracaktır. Son zamanlarda, akademik konuların yazarlara önerilmesine yönelik yapılan birçok çalışma, problemi bağlantı tahmini olarak ele almıştır. Bu çalışmalardan sadece birkaçı, yüksek pratik ilgisine ve üstesinden geldiği belirli zorluklara rağmen iki parçalı ağları kullanmaktadır. İki parçalı ağlarda bağlantı tahmini alanında yapılan önceki çalışmaların çoğu, düğüm çiftleri arasındaki ilişkileri tahmin etmek için iki parçalı ağdan çıkartılan geleneksel tek parçalı ağların özelliklerini kullanmaya odaklanmıştır. Geleneksel projeksiyon yöntemleri, gerçek dünya verilerinden oluşturulan ağlarda kullanıldığında zayıf ilişkilere sahip birçok düğüm çifti içerebilir. Bu zayıf ve gereksiz bilgilerin analizi yüksek hesaplama zamanına neden olur. Bu problemlere bir çözüm olarak, bu çalışmada klasik tek parçalı yöntemlere dönüştürmek yerine ağın omurgası niteliğini taşıyan güçlendirilmiş tek parçalı ağ modeli önerilmiştir. Daha sonra, ağdaki potansiyel bağlantıları tahmin etmek için güçlendirilmiş tek parçalı ağ modeline dayalı benzerlik tabanlı bir tavsiye edici yöntem geliştirilmiştir. Son olarak, önerilen yöntemle elde edilen potansiyel bağlantıların kalitesini değerlendirmek amacıyla zamandan bağımsız ve zaman farkındalı olmak üzere iki yeni yakınlık ölçütü algoritması önerilmiştir. Önerilen yöntemler, IEEE Xplorer çevrimiçi kütüphanesinden toplanan verilerle oluşturduğumuz ağ üzerinden test edilmiştir. Gerçek bir ağ üzerinde yapılan deneyler, önerilen yöntemin başarısının umut verici olduğunu ve büyük ölçekli bir ağdan hızlı ve yüksek kaliteli bağlantı tahmini sonuçları elde etmenin mümkün olduğunu göstermektedir.

Anahtar Kelimeler: İki Parçalı Ağlar, Karmaşık Ağ Analizi, Bağlantı Tahmini, Akademik Konu Önerisi, Güçlendirilmiş Tek Parçalı Ağ, Zaman-farkındalı Yakınlık Ölçütü

SUMMARY

Link Prediction Methods Based on Strengthening Projection Model in Large-Scale Bipartite Networks

Many complex networks created from the real-world data are bipartite structured by nature. Bipartite networks give important ideas about the complex nature of the real-world systems, they are also complex network types that represent the interactions between the different node groups. Recently, the link prediction studies on large-scale and complex networks have particularly become the focus of interest for researchers in various scientific fields. They have a constantly changing and evolving structure. A system that would propose popular topic titles among these networks that are related with an author's fields of work would simplify the work of many researchers. Recently, a lot of work has been done which address recommending academic topics to authors as a link prediction problem. Only a few of these studies use the bipartite networks despite its high practical interest and the specific challenges it raises. The majority of the previous works on the link prediction in bipartite networks focus on using the properties of traditional unimodal projection networks to predict the relations between the node pairs. The traditional projection methods involve many node pairs with weak relationships when used in networks constructed from real-world data. The analysis of these weak and unnecessary information requires high computation time. To cope with this problems, we proposed in this study the notion of strengthening projection networks which is the backbone of the network instead of traditional unimodal projection networks in bipartite graphs. Then, a link prediction method based on strengthening projection network have been developed to predict the potential links. Finally, two new proximity measure algorithms such as time-aware proximity measure algorithm and time-ignored proximity measure algorithm was proposed to evaluate the quality of the potential link. We test the proposed method on the academic information bipartite network constructed with collected data from IEEE XPlorer. Experiments on a real network demonstrate that the success of the proposed method is promising and it is possible to obtain fast and high-quality link prediction results from a large-scale network.

Key Words: Bipartite Networks, Complex Network Analysis, Link Prediction, Academic Topic Recommendation, Strengthening Weighted Projection, Time-aware Proximity Measure

ŞEKİLLER LİSTESİ

Sayfa No

Şekil 2.1.	Karmaşık ağlar.....	5
Şekil 2.2.	Farklı veri kaynaklarından oluşturulan karmaşık ağ örnekleri [35]	6
Şekil 2.3.	Königsberg'in 2 adası ve 7 köprüsü	8
Şekil 2.4.	Bir ağda bulunan farklı kenar türlerine göre gruplandırılmış çizge modelleri	10
Şekil 2.5.	Farklı çizge türleri.....	11
Şekil 2.6.	Düğüm, kenarlar ve kenarların düğüm kümeleri.....	12
Şekil 2.7.	İki parçalı basit bir ağın farklı gösterimleri. İki parçalı ağın iki parçalı çizge ile temsili (a), komşuluk matrisi ile temsili (b).	15
Şekil 2.8.	(a) $G = (U, V, E)$ iki parçalı ağ, (b) U -yansıtılmış tek parçalı ağ, (c) V -yansıtılmış tek parçalı ağ. Şekil 2.8(a)'da üstte halkalarla temsil edilen düğümler U kümesine ait düğümleri, altta kare ile temsil edilen düğümler V kümesine ait düğümleri gösterir.....	17
Şekil 2.9.	Şekil 2.8(a)'da verilen iki parçalı ağdan elde edilmiş U -yansıtılmış ağırlıklı tek parçalı ağ GU (a) ve V -yansıtılmış ağırlıklı tek parçalı ağ GV (b).....	18
Şekil 3.1.	Bağlantı tahmini problemini açıklayan bir örnek [71]	21
Şekil 3.2.	Ağın topolojik özelliklerini kullanan yapısal ve zamansal bağlantı tahmini türleri [83].....	22
Şekil 3.3.	Bağlantı tahmini teknikleri kategorileri.....	24
Şekil 3.4.	İki parçalı ağlarda bağlantı tahmini örneği.....	34
Şekil 3.5:	İç bağlantı tahmini örneği. (a) $G = (U, V, E)$ iki parçalı bir ağ, (b) G iki parçalı ağının U -yansıtılmış $GU = (U, EU)$ tek parçalı ağı, (c) G iki parçalı ağdaki iç bağlantılar, (d) (B, l) iç bağlantısı tarafından oluşturulmuş U -yansıtılmış ağdaki bağlantılar [43].	36
Şekil 3.6.	Kesinlik ve anma arasındaki ilişki [22]	41
Şekil 4.1.	Önerilen yöntemin genel mimarisi	44
Şekil 4.2.	Temel bilgi çıkarma mimarisi.....	46
Şekil 4.3.	İki parçalı yazar-konu ağından bir kesit. (Şekilde, kırmızı renkli olan dairesel düğümler yazarları temsil ederken mavi renkli kare düğümler ise konuları temsil eder).....	47
Şekil 4.4.	(a) Şekil 2.8(b)'de verilen yazar-yazar tek parçalı ağın $\alpha = 1$ için güçlendirilmiş yansıma ağı ($SBPyazar1$), (b) Şekil 2.8(c)'de verilen	

	konu-konu tek parçalı ağın $\alpha = 1$ için güçlendirilmiş yansıma ağı (<i>SBPkonu1</i>).	49
Şekil 4.5.	(a) İki parçalı yazar-konu ağı, (b) İki parçalı yazar-konu eğitim ağı, (c) İki parçalı yazar-konu test ağı.....	52
Şekil 4.6.	İki farklı iki parçalı ağ ve bunların üstteki düğüm kümesinden yansıtılmış ağırlıklı tek parçalı ağ. (b), (a)'da verilen iki parçalı ağdan yansıtılmış tek parçalı ağırlıklı ağ. (d), (c)'de verilen iki parçalı ağdan yansıtılmış tek parçalı ağ.....	53
Şekil 4.7.	İki farklı iki parçalı ağ.	54
Şekil 5.1.	Zaman farkındalı iki parçalı ağın evrimsel gelişimi	66
Şekil 5.2.	Algoritmaların farklı eşik değerlerindeki bağlantı tahmini kalitelerinin AUC değerlendirme ölçütüne göre kıyaslanması	77
Şekil 5.3.	Algoritmaların farklı eşik değerlerindeki bağlantı tahmini kalitelerinin kesinlik değerlendirme ölçütüne göre kıyaslanması	78
Şekil 5.4.	Algoritmaların farklı eşik değerlerindeki bağlantı tahmini kalitelerinin kesinlik- anma eğrileri değerlendirme ölçütüne göre kıyaslanması	79
Şekil 5.5.	$\alpha = 10$ ve $\alpha = 14$ eşik değeriyle güçlendirilmiş daha küçük aralıklı çerçeveli birinci ağ versiyonunda farklı tahmin edicilerin AUC değerleri.....	85
Şekil 5.6.	$\alpha = 10$ ve $\alpha = 14$ eşik değeriyle güçlendirilmiş daha büyük aralıklı çerçeveli ikinci ağ versiyonunda farklı tahmin edicilerin AUC değerleri.....	86

TABLÖLAR LİSTESİ

	<u>Sayfa No</u>
Tablo 3.1. Bağlantı tahmini ölçütlerinde kullanılan ortak semboller	26
Tablo 3.2. Bağlantı tahmini değerlendirme ölçütlerinde kullanılan terimler.....	39
Tablo 4.1. Yıllara göre konferans ve bildiri sayısı.....	45
Tablo 4.2. Eşik değerlerine göre önerilen yöntemle elde edilen potansiyel bağlantı kümelerinin eleman sayıları.....	62



KISALTMALAR

AA	: Adamic/Adat Katsayısı
BP	: Yansıtılmış Ağ
CN	: Ortak Komşular
Cos	: Kosinüs Katsayısı
CT	: Gidip Gelme Zamanı Katsayısı
DN	: Doğru Negatif
DP	: Doğru Pozitif
DPO	: Doğru Pozitif Oranı
HDI	: Hub-Depressed-Index Katsayısı
HPI	: Hub-Promoted-Index Katsayısı
HT	: Ulaşma Zamanı Katsayısı
JK	: Jaccard Katsayısı
LHN	: Leicht-Helman-Nerman Katsayısı
LP	: Yerel Yol Katsayısı
PA	: Tercihli Bağlılık Katsayısı
PL	: Potansiyel Bağlantılar
RA	: Kaynak Paylaşım Katsayısı
RPR	: Köklü PageRank Katsayısı
SAA	: Sosyal Ağ Analizi
Søren	: Sørensen Katsayısı
YN	: Yanlış Negatif
YP	: Yanlış Pozitif
YPO	: Yanlış Pozitif Oranı
LDA	: Levenshtein Mesafe Algoritması

SEMBOLLER LİSTESİ

$\mathbf{A}$	: Komşuluk matrisi
$\mathbf{E}$	: Sonlu kenar kümesi
$\mathbf{U}$	: Sonlu düğüm kümesi
$\mathbf{V}$	: Sonlu düğüm kümesi
$\Gamma(\mathbf{x})$	: x düğümünün komşu dizi seti
$\Gamma(\mathbf{y})$	: y düğümünün komşu dizi seti
$\mathbf{G}$	: İki parçalı ağ
$\mathbf{W}$	: Ağırlık vektörü
r	: Durum değişim oranı
$skor(x, y)$	: x düğümünün y düğümüne bağlanma olasılığı
α	: Uygun eşik değeri
$C_{\omega}(\mathbf{A}, \mathbf{B})$	: ω . çerçevedeki A ve B düğümleri arasındaki tutarlı durum skoru
$D_{\omega}(\mathbf{A}, \mathbf{p})$	: ω . çerçevedeki A ve B düğümleri arasındaki zayıflatıcı durum skoru
$G_{eğitim}$	: İki parçalı eğitim ağı
G_{test}	: İki parçalı test ağı
SBP^{α}	: α ile güçlendirilmiş yansıtılmış ağ
$S_{\omega}(\mathbf{A}, \mathbf{p})$	: ω . çerçevedeki A ve B düğümleri arasındaki güçlendiren durum skoru

1. GİRİŞ

Ağ teorisi, karmaşık sistemlerdeki farklı varlıklar arasındaki etkileşimleri anlamada çeşitli yaklaşımlar sunan disiplinler arası bir bilimdir. Ağlar, hem doğada hem de toplumda bulunan sosyal, biyolojik ve teknolojik bilgi sistemleri gibi çok çeşitli gerçek dünya sistemini temsil etmek amacıyla yaygın bir şekilde kullanılmaktadır. Farklı disiplinlerde çalışan birçok araştırmacının ilgi odağı haline gelen sosyal ağlar oldukça dinamik ve karmaşık bir yapıya sahiptir. Sosyal Ağ Analizi (SAA), bu karmaşıklıkla başa çıkmaya çalışan geniş bir araştırma alanıdır [1]. Sosyal ağlarda zaman içerisinde hem yeni kenarların hem de yeni düğümlerin eklenmesi ya da var olan kenar veya düğümlerin çıkartılmasıyla ağın topolojik yapısında değişimler meydana gelebilir. En önemli sosyal ağ analizi araştırma alanlarından biri olan bağlantı tahmini, ağ içerisindeki bu değişikliklerden bağlantıların durumunda meydana gelen değişikliklerle ilgilenir. Bağlantı tahmini birçok alanda yaygın olarak kullanılmaktadır [2]. Örneğin, bibliyografik alan [3], biyoloji analizi [4], tavsiye amaçlı sistemler [5-7] ve diğer popüler alanlar gibi ağ bilimi ile sınırlı olmayan çok çeşitli alanlarda araştırma yapılmaktadır.

Liben-Nowell ve ark. [7], bağlantı tahmini problemini ve çeşitli benzerlik ölçütlerini ele almıştır. Günümüze kadar bu problemi ele alan birçok yaklaşım sunulmuştur. Literatürde sosyal ağlarda bağlantı tahmini için yaygın olarak kullanılan ölçütler, semantik ve topolojik tabanlı ölçütler olarak kategorize edilebilir [8]. İki düğüm arasındaki ortak komşuların sayısı topolojik ölçütlere bir örnektir. Topolojik ölçütlerden farklı olarak semantik ölçütler, düğümler arası benzerliği hesaplamak için düğümlerin içeriklerini kullanırlar. Her iki benzerlik ölçütü için genel fikir, yüksek benzerlik ölçütlerine sahip olan düğüm çiftlerinin bağlantı kurma olasılığının daha yüksek olmasıdır. Topolojik ölçütler, komşuluk tabanlı ve yol tabanlı ölçütler olarak kategorize edilirler [9]. Komşuluk tabanlı ölçütler, düğümün komşularını göz önüne alır. Bu ölçüte göre iki düğümün ortak komşu sayısı ne kadar fazla olursa düğüm çiftleri arasında gelecekte bağlantı oluşması o kadar muhtemeldir. Yol tabanlı ölçütler ise düğüm çiftleri arasındaki yolları dikkate alır [10]. Bu ölçütün temel mantığı, düğüm çiftleri arasında gelecekte bağlantı oluşma ihtimali aralarındaki yolların kısa olmasıyla doğru orantılıdır. Bununla birlikte, literatürde benimsenen yakınlık ölçütlerinin çoğu ağın evrimsel gelişimini hesaba katmaksızın ağ verisini statik olarak kullanır. Fakat ağın evrimsel gelişimi analiz edildiğinde düğüm çiftleri arasında bazı özel durumların

meydana geldiği görülmüştür. Ağın belirli zaman aralığında bağlı olmayan iki düğüm çifti arasında sonraki aralıkta yeni bir bağlantı oluşabilir veya bağlı olan iki düğüm arasındaki bağ bir sonraki zaman diliminde yok olabilir. Ağa yeni düğümlerin ve bağlantıların eklenmesiyle birlikte düğüm çiftleri arasındaki bağlantı gücü artabilir veya tam tersi azalabilir. Bu durum, ilgili sosyal ağın gelişimini temsil eder. Şüphesiz, sosyal ağın evrimsel gelişimi bağlantı tahmini için yüksek potansiyelli bir bilgi kaynağıdır. Bu sebeple, ağın evrimsel gelişimi bağlantı tahmini sırasında dikkate alınması gereken önemli bir husustur [9].

Sosyal ağ alanında yapılmış çoğu bilimsel çalışma, tek parçalı ağların analizine odaklanır. Fakat gerçek dünya verilerinden oluşturulmuş birçok karmaşık ağ, doğası gereği iki parçalı yapıya sahiptir. Bu sebeple çoğu zaman tek parçalı ağlar gerçek dünya sistemlerini temsil ederken yetersiz kalabilmektedir. İki parçalı ağlar, gerçek dünya sistemlerinin karmaşık yapısı hakkında önemli fikirler vererek farklı düğüm grupları arasındaki etkileşimleri temsil eden özel bir ağ türüdür. Bu ağlar, kenarların yalnızca farklı kümelerdeki düğümler arasında oluştuğu, ancak aynı kümedeki düğümler arasında oluşmadığı, karşılıklı olarak birbirinden bağımsız iki düğüm kümesinden meydana gelir [11,12]. Örneğin a ve b gibi iki ayrı türdeki düğüm kümesinden oluşan iki parçalı bir ağ düşünelim. Örnekte a kümesi aktörleri, b kümesi filmleri temsil etsin ve iki parçalı aktör-film ağı oluşsun. Eğer a düğüm kümesine ait i aktörü, b düğüm kümesine ait j filmde oynamışsa iki parçalı ağda i ve j düğümleri arasında bağlantı vardır, iki parçalı ağdaki aktörlerin ve filmlerin kendi arasında bağlantıları yoktur. Her ne kadar iki parçalı ağlar özel bir amaç için geliştirilmiş ağ türü olarak görünse de kullanım alanları birçok farklı alanda oldukça yaygındır [13]. Son yıllarda, iki parçalı ağlarda bağlantı tahmini [14-16], sosyal ağ analizi [17] ve ilaç yan etki tahmini [18] gibi alanlarda uygulanmaktadır [11].

Son yıllarda, iki parçalı ağların popülerleşmesiyle birlikte, bu alanda yapılan bağlantı tahmini çalışmaları da artmıştır [19-25]. Bu alandaki ilk ve doğal girişim, iki parçalı ağların ayrık düğüm türleri için tek parçalı ağlara dönüştürülmesi ve daha sonra geleneksel tek parçalı ağlarda kullanılan yöntemler ile analiz edilmesidir. Bu yöntemler incelendiğinde iki parçalı ağlardan elde edilen tek parçalı ağlardaki kenar ağırlıklarının özellikle önem taşıdığı görülmüştür. Yansıtılmış ağlardaki bir kenar ağırlığı, bu kenarı oluşturan düğümler arasındaki yakınlık derecesidir. Doğrudan iki parçalı ağlara uygulanabilecek yeni bağlantı tahmini yaklaşımlarının geliştirilmesi, bu yapıların daha iyi anlaşılmasına yönelik önemli bir adımdır.

1.1. Tezin Amacı ve Katkıları

Birçok farklı alan için, bağlantı tahmini, anlaşılması zor ve masraflı olan gelecekteki potansiyel durumu tahmin etmek için kullanılabilir [26-29]. Bu alanlardan birisi de akademik bilgi ağlarıdır. Akademik bilgi ağları, sürekli değişen ve gelişen bir yapıya sahiptir. Bu ağlarda, yazarlara kendi alanlarına uygun popüler konu başlıklarını tavsiye eden bir sistem olması araştırmacıların işini kolaylaştıracaktır. Bu tez çalışmasında akademisyenlere kendi alanlarına uygun çalışabilecekleri konu başlıklarını bulmaları için bir tahmin edici önerilecektir. Bu amaç için öncelikle, yazarlar ve çalıştıkları konu başlıklarını içeren iki parçalı bir akademik bilgi ağı oluşturulacaktır. Oluşturulacak ağ, yazarların akademik çalışmalarındaki konu başlıklarıyla ilişkilendirildiği ağırlıklı bir ağdır. Ağ verisi olarak, bağlantı tahmini çalışmalarında yaygın olarak kullanılan veri kümelerinden farklı olarak IEEE Xplorer çevrimiçi kütüphanesinden amaca yönelik gerçek veri toplanmıştır. Toplanan ham veriden akademik bilgi ağına uygun bilgilerin çıkartılması için yeni bir temel bilgi çıkarım modeli önerilmiştir.

Tez çalışmasında önerilecek yöntem bu alanda yapılan çalışmalardaki birçok probleme çözüm olarak üretilmiştir. Bu tezin literatüre katkıları genel olarak aşağıdaki gibi özetlenebilir:

- Karmaşık ağlar, iki parçalı ağlar ve bağlantı tahmini yöntemleri hakkında temel bilgiler sunulmuştur.
- İki parçalı ağlarda bağlantı tahmini problemlerinin çözümüne yönelik yeni yaklaşımlar geliştirilmiştir.
- Gerçek dünya verilerinden oluşturulmuş iki parçalı karmaşık ağlar analiz edilirken gereksiz bilgi yoğunluğundan dolayı performans düşüklüğü ve çalışma zamanı uzunluğu gibi dezavantajlar ortaya çıkmaktadır. Bu problemlere çare olarak, iki parçalı ağların tek parçalı ağlara dönüştürülmesiyle ilgili güçlendirilmiş tek parçalı ağ modeli olarak adlandırdığımız yeni bir ağ modeli önerilecektir. Önerilen tahmin edici, bu ağ modeline dayalı benzerlik tabanlı bir modeldir.
- İki parçalı ağlar tek parçalı ağlara dönüştürüldüğünde büyük bir bilgi kaybı meydana geldiğinden dolayı önerilen yöntemde hem iki parçalı ağ hem de bu ağdan çıkartılmış güçlendirilmiş tek parçalı ağ bilgileri kullanılmıştır.

- Önerilen yöntemle elde edilen potansiyel bağlantı kümesi için ağın evrimsel gelişiminin hesaba katıldığı ve göz ardı edildiği iki yeni yakınlık ölçütü algoritması önerilmiştir. Önerilen her iki algoritmada da düğüm çiftleri için birden fazla yapısal özellik dikkate alınır.
- Önerilen yöntemler, gerçek hayattan toplanmış özel verilerle oluşturulmuş büyük ölçekli iki parçalı akademik bilgi ağı üzerinde test edilmiştir. Uygulama sonuçları, iki parçalı ağlarda bağlantı tahmini ve iki parçalı ağlarda zaman farkındalı bağlantı tahmini adı altında sunduğumuz her iki yönteminde gerçek dünya verilerinden oluşturulmuş karmaşık ağlardaki başarısını ve uygulanabilirliğini gösterir.

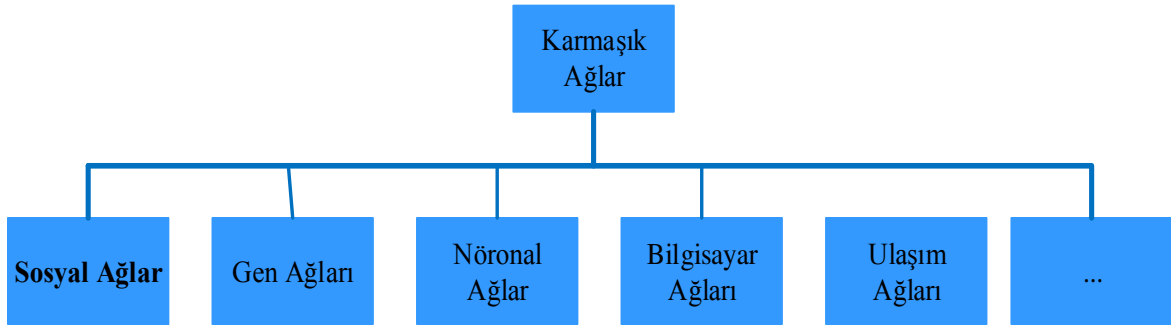
1.2. Tezin Yapısı

Bu tez çalışmasının geri kalan kısımları aşağıdaki gibidir:

- İkinci bölüm karmaşık ağ yapıları ve özellikleri hakkında ayrıntılı bir açıklama içerir. Ayrıca, karmaşık ağ analizlerinde kullanılan temel kavramları sunar. Bu alanda sunulan ağ modelleri sistematik olarak matematiksel tanımlarıyla birlikte sıralanmıştır.
- Üçüncü bölüm, bağlantı tahmini probleminin temel tanımı ve teknikleri hakkında detaylı bilgi sağlar. Bağlantı tahmini için önerilen teknikler; düğüm tabanlı, topolojik tabanlı ve sosyal teori tabanlı olmak üzere üç başlık altında kategorize edilerek detaylandırılmıştır. Ek olarak, iki parçalı ağlarda bağlantı tahmini için sunulan yaklaşımlar hakkında ayrıntılar içerir. Her bir kategori altında geçmişte yapılan çalışmalar hakkında literatür taraması verilmiştir.
- Dördüncü bölümde, iki parçalı akademik bilgi ağı oluşturulmuş ve daha sonra bu ağa uygun tahmin edici model anlatılmıştır. Önerilen tahmin edici için yapılan uygulama sonuçları da bu bölümde verilmiştir.
- Beşinci bölümde, iki parçalı ağın evrimsel gelişimine dayalı zaman farkındalı bağlantı tahmini yaklaşımımız anlatılmıştır. Ayrıca, önerilen tahmin edici için yapılan uygulama sonuçları da yine bu bölümde verilmiştir.
- Altıncı bölümde, bu tezde elde edilen sonuçlar değerlendirilmiş ve ileriye dönük çalışmalar için önerilerde bulunulmuştur.

2. KARMAŞIK AĞLAR

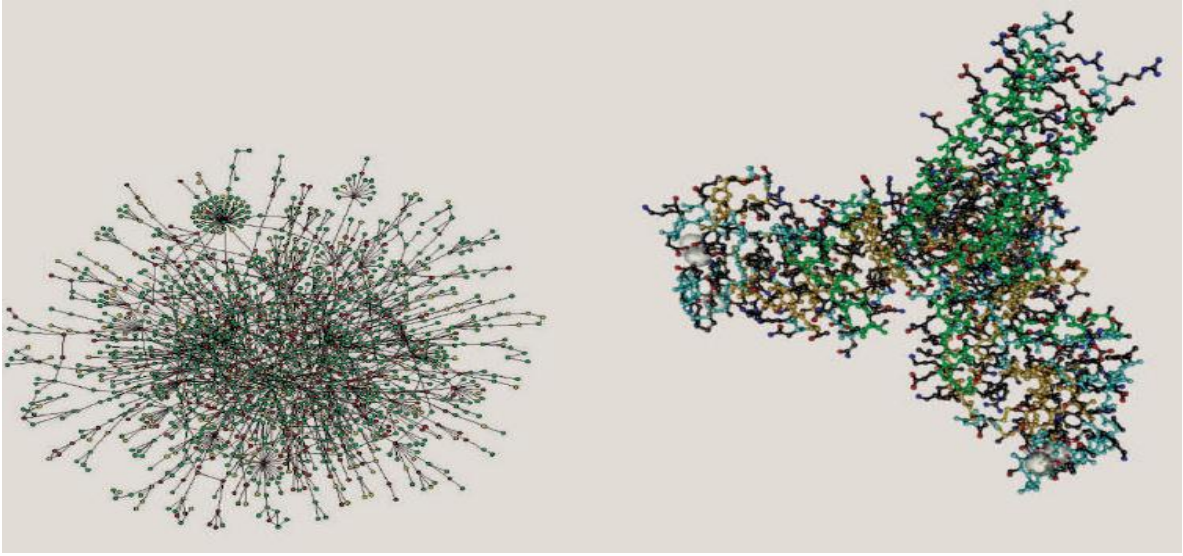
Ağ yapıları, hem doğada hem de toplumda bulunan sosyal, biyolojik ve teknolojik bilgi sistemleri gibi çok çeşitli gerçek dünya sistemini temsil etmek amacıyla yaygın bir şekilde kullanılmaktadır. Bu yapılarda köşeler varlıkları, kenarlar ise köşeler arasındaki ilişkileri veya etkileşimleri temsil eder. Karmaşık bir ağ, gerçek dünya verilerinden oluşturulan realistik bir ağ türüdür. Genellikle gerçek çizgelerde görülen, ancak rastgele çizgeler [30-32] veya ızgara çizgeler [33] gibi basit ağlardaki gibi önemsiz olmayan topolojik özelliklere sahiptir. Çoğu gerçek dünya ağları karmaşık ağlardır. Şekil 2.1’de gerçek dünya karmaşık ağlarına ait bazı örnekler listelenmiştir. Özellikle, arkadaşlık, akrabalık, ortak yazarlık ve ortak ilgi alanları gibi sosyal yapıları modelleyen sosyal ağlar bir tür karmaşık ağ oluştururlar. Karmaşık ağlar ve özellikleri çeşitli alanlardaki araştırmacıların dikkatini çekmiştir. Karmaşık bir ağ, önceden tanımlanmış bir yapıya ya da evrim modeline sahip olmayan soyut bir biçime sahip herhangi bir gerçek dünya ağ olabilir. Sürekli olarak değişen ve gelişen karmaşık ağlar oldukça dinamik yapıya sahiptir. Ayrıca, bu büyük veri çağında küçük bir formla başlayan ağlar kısa sürede çarpıcı büyüklüğe erişebilir. Bu durum, dinamik büyük ölçekli ağları analiz etmek isteyen ağ bilimciler için büyük bir sorun teşkil eder.



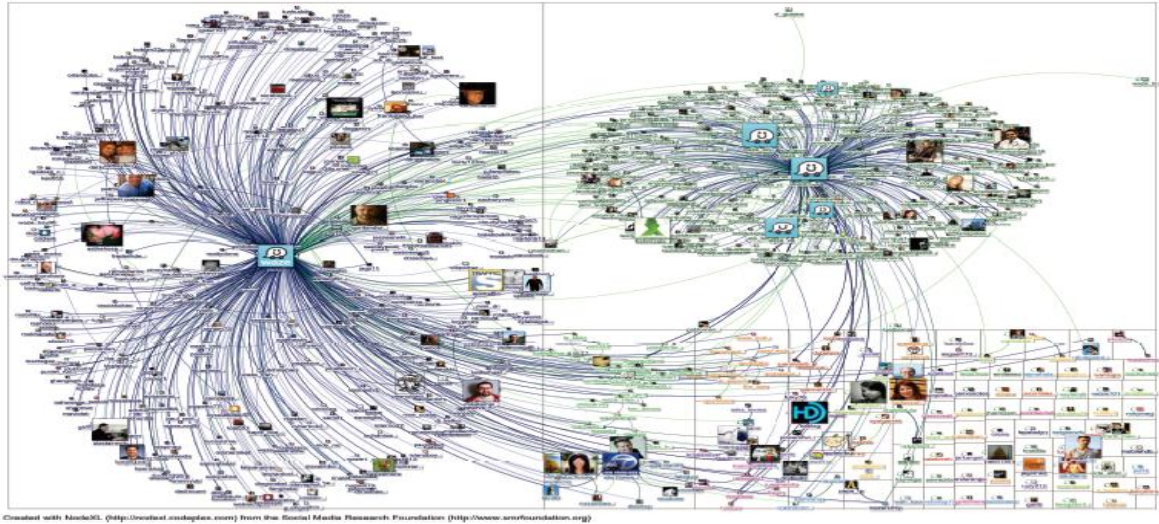
Şekil 2.1. Karmaşık ağlar

Bilgisayar biliminde, internet ve World Wide Web olarak bilinen grafikler şeklinde temsil edilebilen bir dizi gerçek dünya sistemi bulunur. Ayrıca, gıda ağları, protein etkileşimleri, sinir sistemleri veya kan damarları gibi birçok biyolojik ağ da vardır. Yolları veya yerleri birbirine bağlayan ulaşım ağları gibi alt yapı sistemleri ağ yapısı kullanılarak incelenebilir [34]. Günümüzdeki en popüler ağlar, arkadaşlık, güvenlik, akademik, profesyonel veya ticari

işbirliği gibi ilişkileri gösteren sosyal sistemlerdir. Bunun yanı sıra, filmlerin, aktörlerin, çeşitli satın alınan ürünlerin vb. üzerindeki kişilerin gözden geçirilmesini, derecelendirilmesini veya görüşlerini modelleyen ağlar da bulunur. Şekil 2.2, karmaşık ağların iki örneğini gösterir: biri düğümlerin protein olduğu protein-protein etkileşim ağıdır, diğeri ise sosyal paylaşım web sitesi *Twitter*'daki verilerle oluşturulmuş düğümlerin insanları ya da kullanıcıları temsil ettiği, bağlantıların da kullanıcılar arasındaki takipleşme eylemini temsil ettiği arkadaşlık ağıdır.



(a) Protein-protein etkileşim ağı (kaynak <http://plaza.ufl.edu/rkirch05/cis6930/>)



(b) Twitter sosyal etkileşim ağı (kaynak. www.flickr.com/photos/marc_smith/)

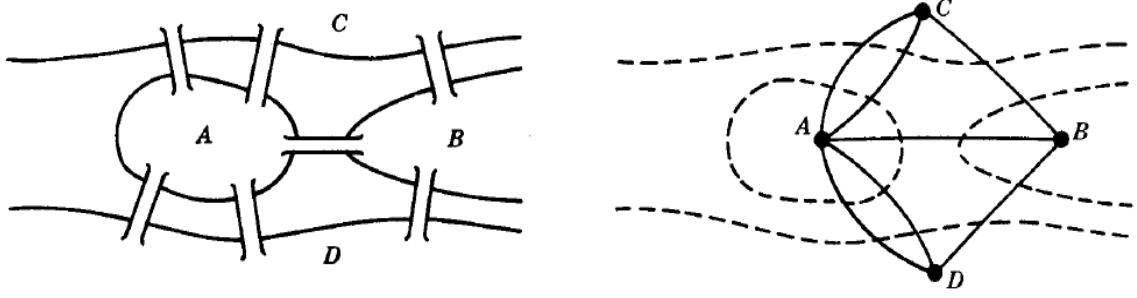
Şekil 2.2. Farklı veri kaynaklarından oluşturulan karmaşık ağ örnekleri [35]

2.1. Karmaşık Ağ Analizi

Geleneksel olarak Sosyal Ağ Analizi (SAA) olarak bilinen karmaşık ağların analizi, ağ yapılarının içerisindeki insanlar, gruplar, kuruluşlar, bilgisayarlar veya diğer varlıklar arası ilişkilerin ve akışların çeşitli teknikler aracılığıyla incelenmesi ve elde edilen verilerden anlamlı sonuçlar türetilmesi olarak tanımlanır. Sosyal ağ analizi ilk zamanlar sosyal ve davranış bilimlerinde daha yoğun bir şekilde kullanılırken günümüzde hemen hemen her alanda kullanılmaktadır [1]. Örneğin, başlangıçta çoğunlukla bireylerin sosyal grup yapılarının ve davranışlarının incelenmesinde kullanılan sosyal ağ analizi teknikleri günümüzde ekonomi, eğitim, ticaret, sağlık ve bankacılık gibi çok daha karmaşık ilişkiler içeren alanlarda da uygulanmaktadır.

Çizge teorisi, karmaşık ağları analiz etmek için önemli bir araç olarak ortaya çıkmıştır. Daha çok sosyoloji ve matematik kökenli olan çizge teorisi; biyoloji, fizik, telekomünikasyon, bilgisayar bilimi vb. alanında ağ analizinde hızlı bir şekilde önem kazanmıştır. Ağın yapısal analizleri, ağların evrimlerini inceleyen zamansal analiz, içerik tabanlı analiz vb. gibi çeşitli analiz bölümleri vardır.

Ağların incelenmesi, 1736 yılında Königsberg köprüsü problemini (Prusya kenti Könisberg köprüsünün her birinden tam olarak bir kez geçecek olan bir yuvarlak yolculuk) çözen İsviçreli matematikçi Leonhard Euler'in matematiksel çizge teorisindeki öncü çalışmasına kadar uzanır. Euler, köprü problemini Şekil 2.3'te gösterildiği gibi çizge ile çözmeye çalışmıştır. 1950-1960'larda, rastgele ağlar (bazı basit rastgele süreçlerle simüle edilen grafikler) Rapoport [36] ve Erdos [37,38] tarafından incelenmiştir. Son yıllarda, modern bilgisayarların ve internetin gelişmesi, büyük ölçekli sosyal ağlar, nöral ağlar ve bilgisayar ağları gibi gerçek dünya karmaşık ağların elde edilmesini, depolanmasını ve analiz edilmesini mümkün hale getirmiştir. Rastgele ağlardan önemli ölçüde farklı olan gerçek ağlarda birçok ilginç ortak özellik ve model keşfedilmiştir [32]. Gerçek ağlarda bulunan ortak ağ özelliklerine örnek olarak küçük dünya ekonomilerini [39], güç kanunu derece dağılımlarını [40], topluluk yapılarını [41], yardımcı karıştırmayı [42] ve sosyal ağların büyüme sürecinde sergilenen diğer kinetik özellikleri verebiliriz. Bu yapılar, insanların dünyamızdaki karmaşık sistemleri anlamalarını büyük ölçüde geliştirmiş ve insanların karmaşık ağ analizlerindeki çok sayıda çalışmaya ilham vermiştir [43].



Şekil 2.3. Königsberg'in 2 adası ve 7 köprüsü

Günümüzde ilgi odağı haline gelen karmaşık ağ analizi antropoloji, sosyoloji, biyoloji, fizik ekonomi, coğrafya ve bilgi bilimleri gibi birçok farklı disiplindeki bilim insanlarının dikkatini çekmiştir. Mevcut araştırma alanları temel olarak gelişmiş kavramların ve ölçümlerin, gerçek veri ve durumlara uygulanmasını, ağ yapısının betimlenmesini, ağın karakteristik yapısının gözlemlenmesini ve ilişkili öğeleri analiz etmek için yeni kavramlar ve ölçümler geliştirilmesini içerir. Ayrıca, ağ yapısının ve işlevinin incelenmesini, ağ topolojisini ve evrimsel gelişimini analiz etmeyi ve sosyal ağ analizi uygulamalarını içerir [44-46].

Temel olarak sosyal ağların analizi aktör, ilişki ve bağ olarak isimlendirilen 3 unsur ile yapılır. Aktörler, bir sosyal ağın en temel unsurudur ve düğümler ile gösterilir. Tüm düğümlerin ilişkileri, bir diyagram ile gösterilir. Bu diyagramda köşeler ve kenarlar, sırasıyla aktör ve ilişkileri gösterir. İlişkiler ise içerik, yön, yoğunluk ve aktif/pasif ilişkiler olmak üzere 4 gruba ayrılır. İçerik, iki aktör arasındaki ilişkiyi gösterir. Yön ise yönlü ve yönsüz olarak sınıflandırılır. Yoğunluk ise zamanla ifade edilir. Buna örnek olarak, bir okulda okuyan öğrenciler arasında iki yıldır süren ilişki verilebilir. Aktif ve pasif ilişkiler ise ilişkinin türünü gösterir. Sosyal ağın analizi için önemli olan bağlar ise güçlü ve zayıf olmak üzere ikiye ayrılır. Güçlü ilişkiler yakın, özel ve aktif ilişkilerdir. Zayıf ilişkiler ise az sayıda ilişki ya da iletişim bilgisi içerirler [47].

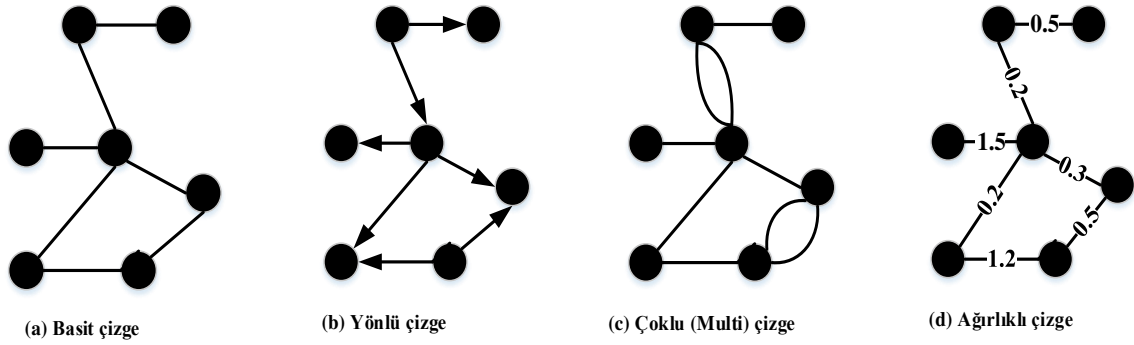
2.2. Çizge Teorisi

Çizge ve ağ terimleri ağ bilimleri literatüründe sıklıkla birbirinin yerine kullanılan iki terimdir. Bu alt bölümde çizge teorisinden alınan terimler, tanımlar ve formüller tanıtılacaktır. Bir çizge diğer bir deyişle ağ, köşeler ve bağlantılar olarak adlandırılan düğümler ve kenarlardan oluşur. Bir çizgenin kenarları, düğüm çiftlerini birbirine bağlar.

Ağların yapısı standart değildir. Düğümler arasındaki belirli bir ilişki türünü belirten bağlantılar yönlü veya yönsüz olabilir. Ayrıca kenarlar ağırlıklı ya da ağırlıksız olabilir. Kenar ağırlığı iki düğüm arasındaki bağlantının kuvvetini temsil eden özelliktir.

Çizgeleri kategorilere ayırmanın farklı yolları vardır. Çizgeler, kenarların içeriği, ağırlıkların varlığı, yönü ve kenar sayısı gibi özelliklere dayalı gruplara ayrılır. Farklı çizge modelleri Şekil 2.4’de gösterilmiştir.

1. **Basit çizgeler**, herhangi bir ağırlık olmaksızın oluşturulan yönsüz, paralel kenar olmayan ve döngü içermeyen (bir düğümden kendisine olan bağlantı) çizge türüdür. Kenarların herhangi bir yönü yoktur ve simetrikler. Düğümler arasındaki ikili ilişkileri temsil ederler.
2. **Yönlü çizgeler**, düğümler arasındaki kenarların yöne sahip olduğu çizge türüdür. Yönlü ağ veya yönlü çizge kısaca digraf olarak adlandırılır. Yönlü ağlara örnek olarak; bir sayfadan diğerine yönlendirilerek çalıştırılan internet ağı, avdan avcıya gelen enerji ile gıda ağı, bir yayından diğerine alıntı yapılan atıf ağı gösterilebilir. Her bir kenarı sıralı bir düğüm çifti ile ilişkilendirilmiştir.
3. **Çoklu (Multi) çizgeler**, basit çizgelerin yeterli olmadığı durumlarda kullanılır. Yönsüz, paralel kenarı olan ve döngü içermeyen çizgelerdir. Basit çizgeler, çoklu çizgedir fakat çoklu çizgeler basit çizge değildir.
4. **Ağırlıklı çizgeler** [48] her bir kenarına nümerik bir değer, ağırlık verilmiş çizge türüdür. Kenar ağırlığı düğümler arasındaki bağlantının gücünü temsil ettiğinden dolayı güçlü ve zayıf olmak üzere ikiye ayrılır. Birçok durumda çoklu grafikler, iki düğümü birbirine bağlayan kenarların sayısı kenar ağırlığına yansıtılacak şekilde ağırlıklı çizgelere dönüştürülür.



Şekil 2.4. Bir ağda bulunan farklı kenar türlerine göre gruplandırılmış çizge modelleri

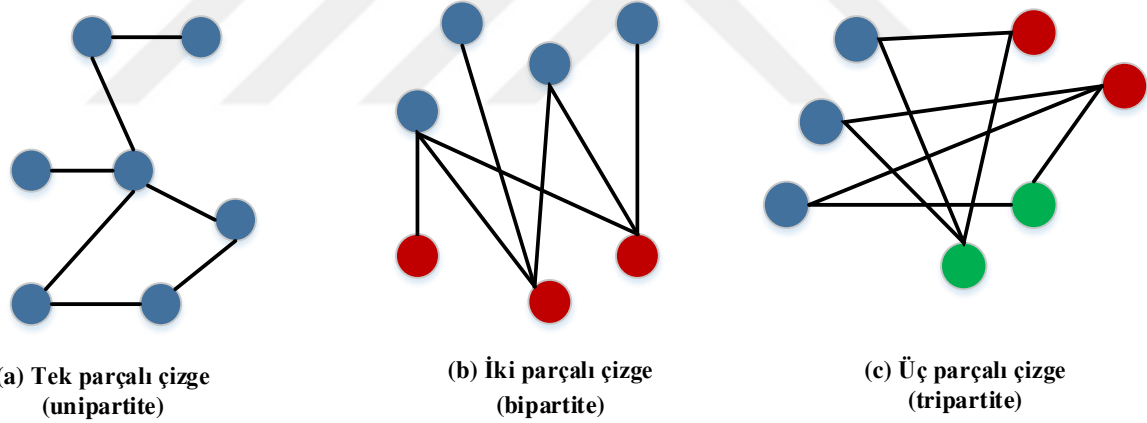
Yönlü ağlarda ilişkiler tek yönlü ya da çift yönlü olabilir. Örneğin Twitter’da bir kişi ikinci bir kişiyi takip ettiğinde ya da ona yorum yazdığında kurulan ilişki tek yönlüdür. Eğer ikinci kişi, ilkinin cevap yazar ya da takip ederse ilişki çift yönlüye dönüşür. Ağ grafiklerinde tek yönlü ilişkiler tek taraflı okla, çift yönlü ilişkiler çift taraflı okla, yönsüz ilişkiler ise oksuz çizgilerle gösterilir.

Çizgelerin için bir başka ayrım, düğümler arasındaki bağlantıların türüne göre yapılabilir. Şekil 2.5’de verilen çizgeler bağlantı türüne göre kategorileştirilmiş çizge modellerine örnektir. Çizge içerisindeki bir düğümün bağlantısı ya hiçbir zaman kendi kümesindeki düğümlerle oluşmayıp ancak farklı kümelerdeki düğümlerle oluşabilir, ya da yalnızca kendi kümesindeki düğümlerle oluşabilir. Bu kavramlar, çizgeleri aşağıdaki gibi kategorize etmemizi sağlar.

- 1. Tek parçalı (Unipartite) çizgeler:** Herhangi bir düğüm bölümünün olmadığı çizge türüdür. Yalnızca bir düğüm kümesi vardır ve her düğüm bir başka düğüme bağlı olabilir. Örneğin, ortak yazarlık veya bilimsel işbirliği ağı tek parçalı ağıdır.
- 2. İki parçalı (Bipartite) çizgeler:** İki parçalı bir çizgenin iki düğüm kümesi vardır ve kenarlar yalnızca farklı düğüm kümeleri arasında oluşur. Aynı düğüm kümesindeki düğümler arasında hiçbir bağlantı yoktur. Örneğin; yazarlar ve yaptıkları bilimsel makalelerden oluşan yazar-makale akademik bilgi ağı iki parçalı bir ağıdır. Diğer bir örnek olarak, öğrencilerin aldıkları dersleri temsil etmek amacıyla oluşturulan bir taraflı öğrenciler diğer taraflı aldıkları dersler olan öğrenci-ders ağı da iki parçalı bir ağıdır. Bu ağlarda bağlantılar, yalnızca yazarlar ve bilimsel makale yayınları arasında oluşur. İki parçalı ağlar tek parçalı ağlara dönüştürülebilir. Örneğin, yazar-makale iki parçalı ağı,

sadece yazarlardan ve sadece bilimsel makalelerden oluşan tek parçalı iki ağa ayrıştırılabilir. Tek parçalı yazar-yazar ağı oluşturulurken iki parçalı ağda en az bir ortak makaleye sahip yazarlar birbirine bağlanır. Benzer şekilde, tek parçalı makale-makale ağı oluşturulurken iki parçalı ağda en az bir ortak yazara sahip olan makaleler birbirine bağlanır. Tek parçalı çizge modeli çok kullanışlı bir yöntemdir. Fakat ikili parçalı ağların tek parçalı ağlara dönüştürülmesi ağ ile ilgili birçok bilginin göz ardı edilmesine sebep olur. Örneğin, yazar-makale ağı tek parçalı ağa dönüştürüldüğünde hangi yazarların hangi makaleye katkıda bulunduğu bilgisi iki parçalı ağda mevcutken tek parçalı yazar-yazar ağında bu bilgiye erişilemez. İki parçalı ağların ve tek parçalı ağlara dönüştürülmeleri Bölüm 2.5’te detaylı olarak verilmiştir.

3. **Üç parçalı (Tripartite) çizgeler:** Üç farklı düğüm kümesinden oluşan çizge türüdür. Bu çizge türünde de bağlantılar sadece farklı kümelerle ait düğümler arasında meydana gelir.



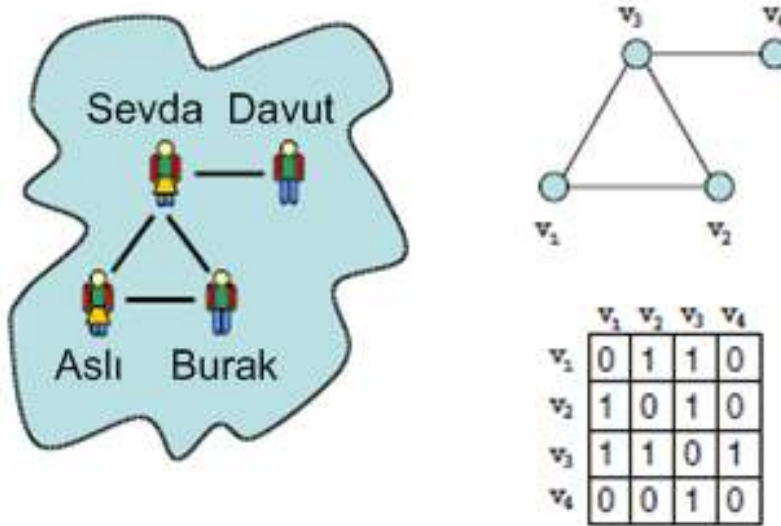
Şekil 2.5. Farklı çizge türleri

2.2.1. Matematiksel Tanımlar

Matematiksel olarak basit bir çizge $G = (V, E)$ ile temsil edilir. Burada $V = \{v_1, v_2, \dots, v_n\}$, $|V| = n$ ağdaki sonlu düğüm kümesini diğer bir deyişle köşe kümesini temsil ederken $E \subseteq V \times V$, $E = (v_i, v_j)$, $i \neq j$, $|E| = m$ ise sonlu bağlantı kümesini diğer bir deyişle kenar kümesini temsil eder. Her bir kenar kendisi ile ilişkilendirmiş bir ya da iki adet düğüm içerir ve bu düğümleri birbirine bağlar. Bir G çizgesi $n \times n$ boyutunda A komşuluk matrisiyle ifade edilir. a_{ij} , A komşuluk matrisinin bir elemanı olmak üzere düğümler arasındaki bağlantının varlığına veya yokluğuna bağlı olarak 0 veya 1 değerini alır.

$$a_{ij} = \begin{cases} 0 & \text{eğer } i \text{ ve } j \text{ düğümleri arasında bir kenar varsa } ((i, j) \notin E), \\ 1 & \text{eğer } i \text{ ve } j \text{ düğümleri arasında bir kenar yoksa } ((i, j) \in E) \end{cases}$$

Şekil 2.6, basit bir çizge modelini ve bunun komşuluk matrisini temsil eder. Şekil 2.6'da da görüldüğü gibi E kümesindeki her bir bağlantı V kümesindeki herhangi iki düğüm arasındaki bağlantıya eşdeğerdir. Örneğin, v_1 ve v_2 uç noktaları e_1 kenarı ile bağlanmıştır.



Şekil 2.6. Düğümler, kenarlar ve kenarların düğüm kümeleri

Çizge teorisinin temelleri ilk olarak [49] numaralı çalışmada sunulmuştur. Sonraları, [50,51] numaralı çalışmalarda daha kapsamlı bir şekilde incelenmiştir.

2.3. Karmaşık Ağların Özellikleri

Neredeyse tüm karmaşık ağ türleri bazı ortak yapısal özelliklere sahiptir. Bunların en önemlileri aşağıda sıralanmıştır.

Derece Merkeziliği: Bir ağdaki en basit ama en aydınlatıcı merkezilik ölçütüdür. Klasik olarak bir düğümün ağ içerisinde sahip olduğu kenar sayısını temsil eder. Sosyal ağlarda, diğer birçok kişi ile bağlantısı olan bireylerin daha fazla etkiye sahip olduğunu, bilgiye daha rahat eriştiğini veya daha az bağlantıya sahip olan bireylerden daha prestijli olabileceğini söylemek mümkündür. Yönlü ağlarda köşeler, hem giriş derecesi hem de çıkış derecesine sahiptir ve ikisi de uygun koşullarda merkezilik ölçütü olarak yararlı olabilir [52].

Yakınlık merkeziliği: Ağdaki herhangi bir düğümden diğer tüm düğümlere olan en kısa yolların ortalama uzaklıklarının (geodesic distance) toplanmasıyla elde edilir. Bu ölçüm sosyal ağdaki bileşen sayısına bağlıdır. Merkeziliğin tamamen farklı bir ölçütü, bir köşeden diğerine olan uzaklık anlamına gelen yakınlık merkeziliği ile sağlanmaktadır. Dolayısıyla bu ölçüm farklı büyüklükteki sosyal ağlar arasında karşılaştırmaya uygun değildir.

Arasındalık merkeziliği: Ağdaki bir düğümden ya da aktörden geçen en kısa yolların oranı ile bulunur. Büyüklüğü fazla ağlarda hesaplanması çok maliyetli olabilecek bir ölçüt olduğu için belirli seviyedeki komşulara kadar inilerekte hesaplanabilir. Arasındalık derecesi yüksek olan aktörler konumları dolayısıyla, diğer aktörlere göre daha önemli bir konumdadır ve ağda olup bitenden daha çok haberdardır.

Köprü: Bağlantının kaldırılması diğer uç noktalara (düğümlere) ulaşamamasına sebep oluyorsa bu bağlantıya köprü diyebiliriz. Köprü, ağ içerisindeki grupları bağlar ve kaldırıldığında gruplar arası bağlantıda ortadan kaybolmuş olur.

Yol Uzunluğu: Ağ içerisindeki düğümler arası mesafe olarak tanımlanır. Düğümler arası mesafe ne kadar kısaysa aralarında o kadar az sayıda düğüm vardır. Düğümlerin ilişki kurma ihtimali aralarındaki mesafeyle ters orantılıdır.

Kümelenme katsayısı: Bir düğümün komşularıyla olan bağlantı gücüdür. Bir düğümün komşularıyla arasındaki bağlantı gücü gerçek bağlantıların olası bağlantılara oranıyla hesaplanır.

Bağlılık: Ağdaki düğümler arası bağlantıların gücünü ölçer. Bağlantı gücü ne kadar yüksekse düğümlerin birbirlerine olan benzerliği de o kadar fazladır.

Yoğunluk: Mevcut ağda var olan tüm bağlantıların, oluşması muhtemel bağlantılara oranıyla hesaplanır.

2.4. İki Parçalı Ağlar

İki parçalı ağlar, gerçek dünya sistemlerinin karmaşık yapısını temsil eden özel bir ağ türüdür. Bu ağ türü, klasik tek parçalı ağlardan farklı olarak iki farklı düğüm kümesine ayrılabilen ve aynı düğüm kümesindeki hiçbir düğümün komşu olmadığı bir yapıya sahiptir [11,12]. Birkaç örnek vermek gerekirse, aktörlerin oynadıkları filmlerle bağlantılı olduğu aktör-film ağı [53], yazarların birlikte yaptığı çalışmalarda ortak yazarlık ağı ve atıf ağları [54,55], hastalık gen ağı [56-59], RNA-protein etkileşimi [60], müşterilerin aldıkları ürünlere bağlı olduğu çevrimiçi alışveriş ağları [61] vb. verilebilir.

İki parçalı ağ modelleri, kenarların yalnızca farklı kümelerdeki düğümler arasında oluştuğu, ancak aynı kümedeki düğümler arasında oluşmadığı, karşılıklı olarak birbirinden bağımsız iki düğüm kümesinden oluşur. İki parçalı bir ağ, $G = (U, V, E)$ şeklinde tanımlanır. Burada, U ve V ayrık düğüm kümeleridir. E kümesi ise G iki parçalı ağın bir kenar kümesi, $U \times V$ 'nin alt kümesidir. $\Gamma(U)$, G ağında U düğümlerinin komşuluk kümesi, $\Gamma(V)$ ise G ağında V düğümlerinin komşuluk kümesidir. İki parçalı ağlarda $(u, v) \in E$ olan tüm bağlantılarda $u \in U$ ve $v \in V$ ya da $v \in U$ ve $u \in V$ 'dir.

Matematiksel olarak ifade etmek gerekirse:

Tanım 1 (İki Parçalı Ağ): İki parçalı bir ağ U ve V iki düğüm kümesi ve $E \subseteq |U| \times |V|$ kenar kümesinden oluşan $G = (U, V, E)$ üçlüsüdür. Aynı, U ve V kümeleri arasında kenar yoktur.

Genellikle, iki parçalı bir ağ $|U| \times |V|$ matrisiyle temsil edilir. G ağının U tarafında n tane düğüm, V tarafında m tane düğüm olduğunu varsayarsak, G iki parçalı ağ $m \times n$ boyutunda

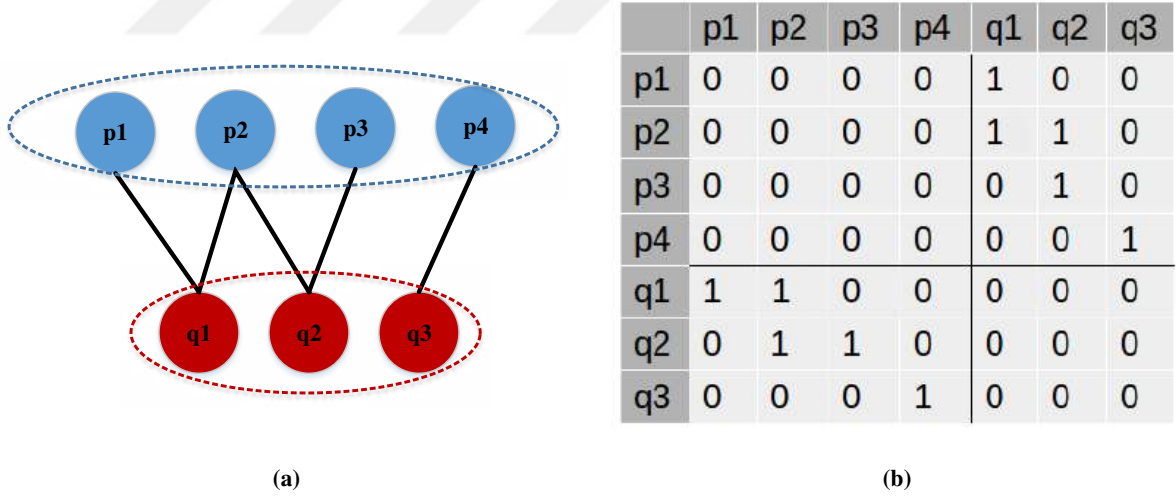
komşuluk matrisiyle temsil edilebilir [62]. G iki parçalı ağın matris içeriği ve A komşuluk matrisini temsil eden köşegen matris aşağıdaki gibidir:

$$a_{ij} = \begin{cases} 1, & \text{eğer } (u_i, v_j) \in E, \\ 0, & \text{aksi taktirde.} \end{cases} \quad (2.1)$$

$$A = \begin{bmatrix} 0_{n \times n} & A_{n \times m} \\ A_{m \times n}^T & 0_{m \times m} \end{bmatrix}. \quad (2.2)$$

Burada $0_{n \times n}$ ve $0_{m \times m}$ sırasıyla $n \times n$ ve $m \times m$ tüm sıfır matrislerdir. $A_{n \times m}$ ve $A_{m \times n}$, sırasıyla $n \times m$ ve $m \times n$ sıfır olmayan bir matristir. Komşuluk matrisi simetrik olduğundan, iki parçalı G ağını temsil etmek için sadece $A_{n \times m}$ matrisini kullanabiliriz; her bir satır U kümesindeki ve her sütun V kümesindeki bir düğümü temsil etmektedir [11].

Şekil 2.7’de iki parçalı basit bir ağ ve bu ağın komşuluk matrisi verilmiştir. Şekilde açıkça görülebileceği gibi iki parçalı ağ ayrık iki düğüm kümesinden oluşur.



Şekil 2.7. İki parçalı basit bir ağın farklı gösterimleri. İki parçalı ağın iki parçalı çizge ile temsili (a), komşuluk matrisi ile temsili (b).

2.4.1. İki Parçalı Ağların Tek Parçalı Ağlara Yansıtılması

Çoğu ağ analiz ölçütleri doğrudan iki parçalı ağlara uygulanamaz. Bu nedenle, iki parçalı ağ analizinde kullanılan en yaygın yöntemlerden biri, orijinal iki parçalı ağa kodlanmış bilgileri mümkün olduğunca koruyan tek parçalı ağ modellerine dönüştürülmesidir. İki parçalı ağlar incelenirken genellikle “Tek-Modlu Ağ” olarak adlandırılan tek parçalı ağlara dönüştürülerek incelenir [63]. Tek parçalı ağlara dönüştürmek aslında iki parçalı ağı kendisini oluşturan ayırık kümelerden birine indirgemektir. Yansıtılmış diğer bir deyişle izdüşümü alınmış bir ağ, bilinen klasik tek parçalı ağ yapısıdır. İki parçalı ağlar tek parçalı ağlara dönüştürülürken ağın topolojik yapısında gizli bazı bilgiler önemini kaybedebilir. Bu sebeple, sadece yansıtılmış ağların kullanıldığı iki parçalı ağ analizi çalışmalarının sonuçları yetersiz olabilmektedir.

Matematiksel olarak ifade etmek gerekirse:

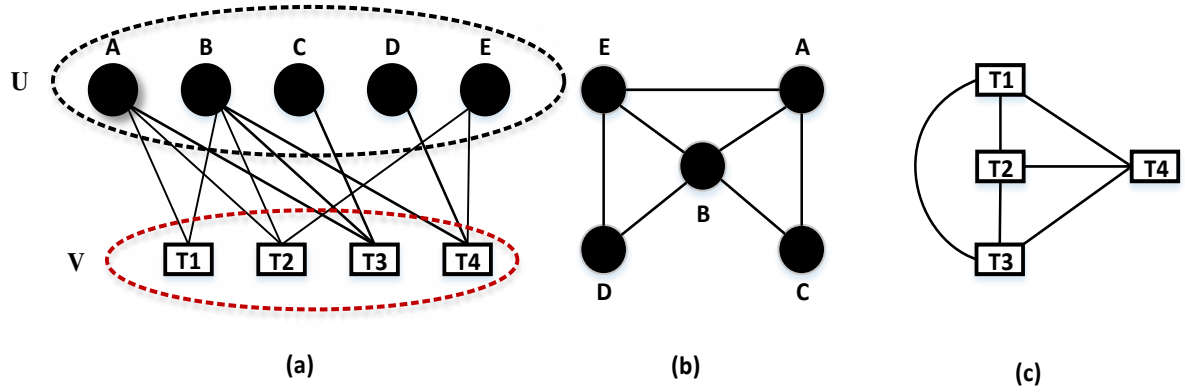
Tanım 2 (Yansıtılmış Ağ): $G = (U, V, E)$, $|U| = m$, $|V| = n$ ve $E \subseteq U \times V$ boyutlarında iki parçalı bir ağdır. İki parçalı ağ kendini oluşturan her bir U ve V düğüm kümelerinden, U -yansıtılmış ağ $G_U = (U, E_U)$ ve V -yansıtılmış ağ $G_V = (V, E_V)$ olmak üzere iki tane tek parçalı ağa dönüştürülebilir. Bunun için:

$$E_U = \{(u_i, u_j) \mid u_i, u_j \in U, \exists v_i \in V, v_i \in \Gamma(u_i) \cap \Gamma(u_j)\}, \quad (2.3)$$

$$E_V = \{(v_i, v_j) \mid v_i, v_j \in V, \exists u_i \in U, u_i \in \Gamma(v_i) \cap \Gamma(v_j)\}. \quad (2.4)$$

Tanımdan da anlaşıldığı gibi, iki parçalı ağda U düğüm kümesinin elemanlarından olan u_i ve u_j düğümleri, iki parçalı ağda V düğüm kümesinde en az bir tane ortak komşuya sahiplerse, G_U yansıtılmış ağda (u_i, u_j) düğümleri arasında bağlantı vardır.

Benzer şekilde, G 'nin V -yansıtılmış ağı, $G_V = (V, E_V)$ şeklinde tanımlanır. Örneğin, Şekil 2.8 üstteki daire şeklindeki düğümlerin yazarları, alttaki kare düğümlerin çalışma konularını temsil ettiği, iki parçalı yazar-konu ağıdır. Şekil 2.8(a)'da gösterilen G iki parçalı bir sosyal ağdır, sırasıyla Şekil 2.8(b) ve Şekil 2.8(c) izdüşümü alınan G iki parçalı ağın, U -yansıtılmış ve V -yansıtılmış tek parçalı ağ eşlenikleridir.



Şekil 2.8. (a) $G = (U, V, E)$ iki parçalı ağ, (b) U -yansıtılmış tek parçalı ağ, (c) V - yansıtılmış tek parçalı ağ. Şekil 2.8(a)'da üstte halkalarla temsil edilen düğümler U kümesine ait düğümleri, altta kare ile temsil edilen düğümler V kümesine ait düğümleri gösterir.

2.4.2. İki Parçalı Ağların Ağırlıklı Tek Parçalı Ağlara Yansıtılması

İki parçalı ağları tek parçalı ağlara dönüştürürken ortaya çıkan birçok soruna rağmen, araştırmacılar iki parçalı ağ analizlerinde tek parçalı ağlara dönüştürme tekniğini kullanmaktadır. İki parçalı ağlar tek parçalı ağlara dönüştürülürken ağın topolojik yapısında gizlenmiş bazı bilgiler önemini kaybedebilir. Bir nebze de olsa bu durumun önüne geçebilmek için ağırlıklı iki parçalı ağlar kullanılır. Ağırlıklı ağlar birçok alanda yaygındır, fakat yazar-makale ilişkileri, hasta-hastalık, aktör-film gibi iki parçalı ağlarda ağın topolojik özelliklerini temsil ederken özellikle önemlidir. Yazarlar ve yaptıkları yayınları içeren iki parçalı ağdan çıkartılan tek parçalı yazar-yazar ağında bir kenar ağırlığı iki yazarın iki parçalı ağda birlikte yaptığı yayın sayısını temsil eder. Ağırlıklı tek parçalı ağ yansıması, iki boyutlu bir ağdan ağırlıklı tek parçalı bir ağ elde etmek için kullanılan bir tekniktir; buradaki kenar ağırlıkları aksi belirtilmedikçe köşelerin ortak komşu sayısını temsil eder ve kenar ağırlıkları kullanılarak elde edilen ağa ağırlıklı yansıtılmış ağ adı verilir.

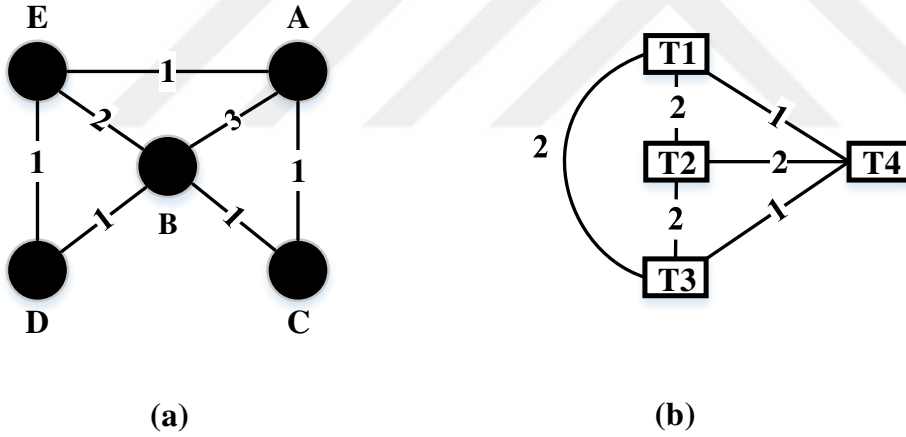
Matematiksel olarak ifade etmek gerekirse:

Tanım 3 (Ağırlıklı Yansıtılmış Ağ (BP)): $G = (U, V, E)$, $|U| = m$, $|V| = n$ ve $E \subseteq U \times V$ boyutlarında iki parçalı bir ağdır. İki parçalı ağ kendini oluşturan her bir U ve V düğüm kümelerinden, U -yansıtılmış ağ $G_U = (U, E_U)$ ve V -yansıtılmış $G_V = (V, E_V)$ olmak üzere iki tane tek parçalı ağa dönüştürülebilir. Elde edilen tek parçalı ağlarda her bir E_U ve E_V kenarları için W ağırlık fonksiyonu kullanılarak ağırlıklı tek parçalı ağ elde edilir.

$$E_U = \{(u_i, u_j) \mid u_i, u_j \in U, \exists v_i \in V, v_i \in \Gamma(u_i) \cap \Gamma(u_j)\}, W: (u_i, u_j) \rightarrow |\Gamma(u_i) \cap \Gamma(u_j)|, \quad (2.5)$$

$$E_V = \{(v_i, v_j) \mid v_i, v_j \in V, \exists u_i \in U, u_i \in \Gamma(v_i) \cap \Gamma(v_j)\}, W: (v_i, v_j) \rightarrow |\Gamma(v_i) \cap \Gamma(v_j)|. \quad (2.6)$$

Genelliği kaybetmeksizin $W(u_i, u_j)$, u_i düğümünün u_j düğümü ile olan önemini temsil eder. U tek-parçalı ağ için $W(u_i, u_j)$, $\Gamma(i) \cap \Gamma(j)$ ortak komşu sayısıdır. Örneğin Şekil 2.8(a)'da verilen iki parçalı ağda $A \in U$ için $\Gamma(A) = \{T1, T2, T3\}$, A düğümünün iki parçalı ağdaki komşu sayısı; $B \in U$ için $\Gamma(B) = \{T1, T2, T3, T4\}$, B düğümünün iki parçalı ağdaki komşu sayısıdır. Bu iki kenarın $\Gamma(A) \cap \Gamma(B) = \{T1, T2, T3\}$ ortak komşu sayısı üç olduğundan Şekil 2.9(a)'da ifade edilen $G_U = (U, E_U)$, U - yansıtılmış ağırlıklı tek parçalı ağda $W(A, B)=3$ 'tür.



Şekil 2.9. Şekil 2.8(a)'da verilen iki parçalı ağdan elde edilmiş U -yansıtılmış ağırlıklı tek parçalı ağ G_U (a) ve V -yansıtılmış ağırlıklı tek parçalı ağ G_V (b).

3. BAĞLANTI TAHMİNİ

Sosyal ağlar aktörler ve aktörler arası ilişkilerden oluşan karmaşık bilgi ağlarıdır. Bu ağlar aynı ya da farklı tipteki kişilerin ya da diğer çokluklar arasındaki etkileşimi, işbirliğini gösteren bir yapıdır. Sosyal ağlarda bulunan eleman ya da varlıklar düğümleri oluştururken düğümler arasındaki etkileşimler ise bağlantıları temsil etmektedir. Sosyal ağlar doğası gereği oldukça karmaşık ve dinamik bir yapıya sahiptir. Ağ içerisinde yeni bağlantılar ve varlıklar zamanla görülür veya kaybolur. Bu durum, onları büyük ölçüde dinamik ve karmaşık sistemler yapar. Bağlantı tahmini, karmaşık ağlardaki komşu olmayan düğümler arasındaki belirsizlik ve potansiyel ilişkiyi analiz etmek için kullanılan etkili bir matematiksel araçtır [64]. Bağlantı tahmini, sosyal ağ içerisindeki değişimlerden iki düğüm arasındaki bağlantılara odaklanan etkili bir sosyal ağ analizi aracıdır. Sosyal ağlarda gelecekte yeni ilişkilerin oluşacağını 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 [65].

Bağlantı tahmini, anlaşılması doğrudan zor olan muhtemel ilişkileri ortaya çıkarmak veya gelecek davranışı tahmin etmek için kullanılır [66]. Bağlantı tahmini yöntemleri sosyal, biyolojik ve bilgi sistemleri gibi farklı disiplinlerde çalışan birçok araştırmacının dikkatini çekmiştir. Günümüzde bibliyografik alan [3], moleküler biyoloji [40], adli soruşturmalar [67], tavsiye amaçlı sistemler [68], tıp [69,70] vb. alanlarda bağlantı tahmini yapısı kullanılmaktadır. Bu işlem birçok çalışma içerisinde farklı amaçlar için kullanılmaktadır. Sosyal ağlarda gelecekte, hangi düğümler arasında yeni ilişkiler olabilir veya var olan hangi ilişkiler kaybolabilir, düğümler arası ilişkilerin yoğunluğuna bağlı olarak ağda hangi gruplar vardır, hangi düğümler hangi gruptalardır, hangi düğümler diğer düğümleri daha çok etkilemektedir vb. soruların cevapları araştırılmaktadır. Son yıllarda, bu alanda yapılan çalışmalar büyük bir artış sergilemektedir.

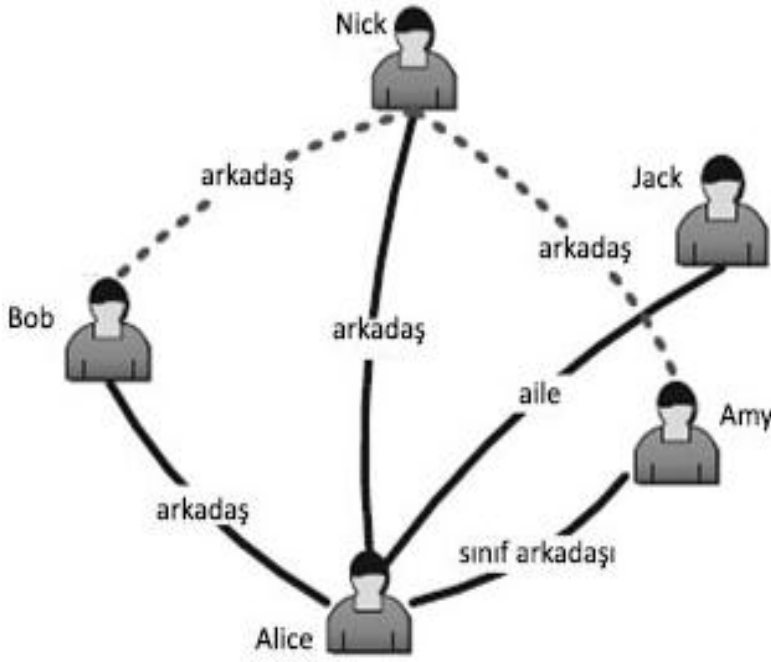
Bağlantı tahmini birçok önemli alanda uygulanabilir. Büyük organizasyonlara, oluşacak iş birliktelikleri tavsiye edilebilir, akademik yazarlara çalışabilecekleri yazarlar ya da konular önerilebilir, sabıkalıların sosyal medyadaki ya da telefon görüşmelerindeki ilişkiler

incelenerek oluşabilecek suç eylemleri önceden sezilebilir, web sitelerinde bağlantı tahmini yapılabilir, e-ticarete kullanıcıların davranışlarına göre en uygun ürünler tavsiye edilebilir, biyoinformatikte protein-protein, hastalık-gen, hastalık-ilaç etkileşimleri tahmin edilebilir, herhangi bir tamamlanmamış sosyal ağ verisinin eksik kısmı tahmin edilebilir. Örneğin; arkadaşlık ağlarında bireylere yeni arkadaşların önerilmesi, hastalık ağlarında bağlantılı hastalıklardan yola çıkarak hastalık risklerinin tespit edilmesi, ortak yazarlık ağlarında yazarlara birlikte çalışma yapabilecekleri yazarların önerilmesi, yazar-makale ağlarında yazarlara çalışabilecekleri konuların önerilmesi bağlantı tahmininin kullanıldığı alanlardan birkaçıdır.

3.1. Bağlantı Tahmini Problemi

Yeni bağlantıların tahmin edilmesi problemi (ya da basitçe bağlantı tahmini problemi) belirli bir zaman periyodu boyunca ağdaki bağlantıların ortaya çıkma ve kaybolma geçmişini inceleyerek ağdaki bağlantıların gelecekteki yapısıyla ilgili önsezilerde bulunulmasıdır. Sosyal ağ analizinin en önemli araştırma alanlarından biri olan bağlantı tahmini problemi klasik olarak şöyle tanımlanmıştır: “Bağlantı tahmini t zamanında verilen bir sosyal ağın anlık durumu göz önünde bulundurulduğunda, t zaman aralığından belirli gelecek zamana kadar ağa eklenecek olan kenarları doğru tahmin etme işlemidir.” [7].

Matematiksel olarak tanımlamak gerekirse, belirli bir t zamanında, sırasıyla V 'nin düğüm kümesini, E 'nin bağlantı kümesini temsil ettiği $G = (V, E)$ sosyal ağını düşünelim. Bağlantı tahmini, $t'(t' > t)$ gelecek zamanı için düğümler arasında kaybolmuş veya yeni oluşmuş bağlantıları ya da mevcut ağda eksik ya da gözlemlenmemiş bağlantıları tahmin etmeyi hedefler. Bağlantı tahmini problemi, Şekil 3.1’de verilen beş kişinin arkadaşlık ilişkilerinin temsil edildiği sosyal ağla açıklayalım. Burada kalın ve kesiksiz çizgiler t zamanında varolan etkileşimleri gösterirken kesiksiz çizgilerde $[t, t']$ zaman aralığında yeni ortaya çıkmış etkileşimleri gösterir. t zamanında, Alice ve Bob arkadaşdır, Alice aynı zamanda Nick ile de arkadaşdır. Belki de t' zamanında Alice, Nick ile Bob’u tanıştırdı ve böylelikle arkadaş oldular. Benzer şekilde, Nick ve Amy’de t' zamanında arkadaş olurlar. Burada bağlantı tahmini probleminin amacı, kişilerin arasında yeni oluşacak arkadaşlık ilişkilerinin tahmin edilmesidir.



Şekil 3.1. Bağlantı tahmini problemi açıklayan bir örnek [71]

3.2. Bağlantı Tahmini Teknikleri

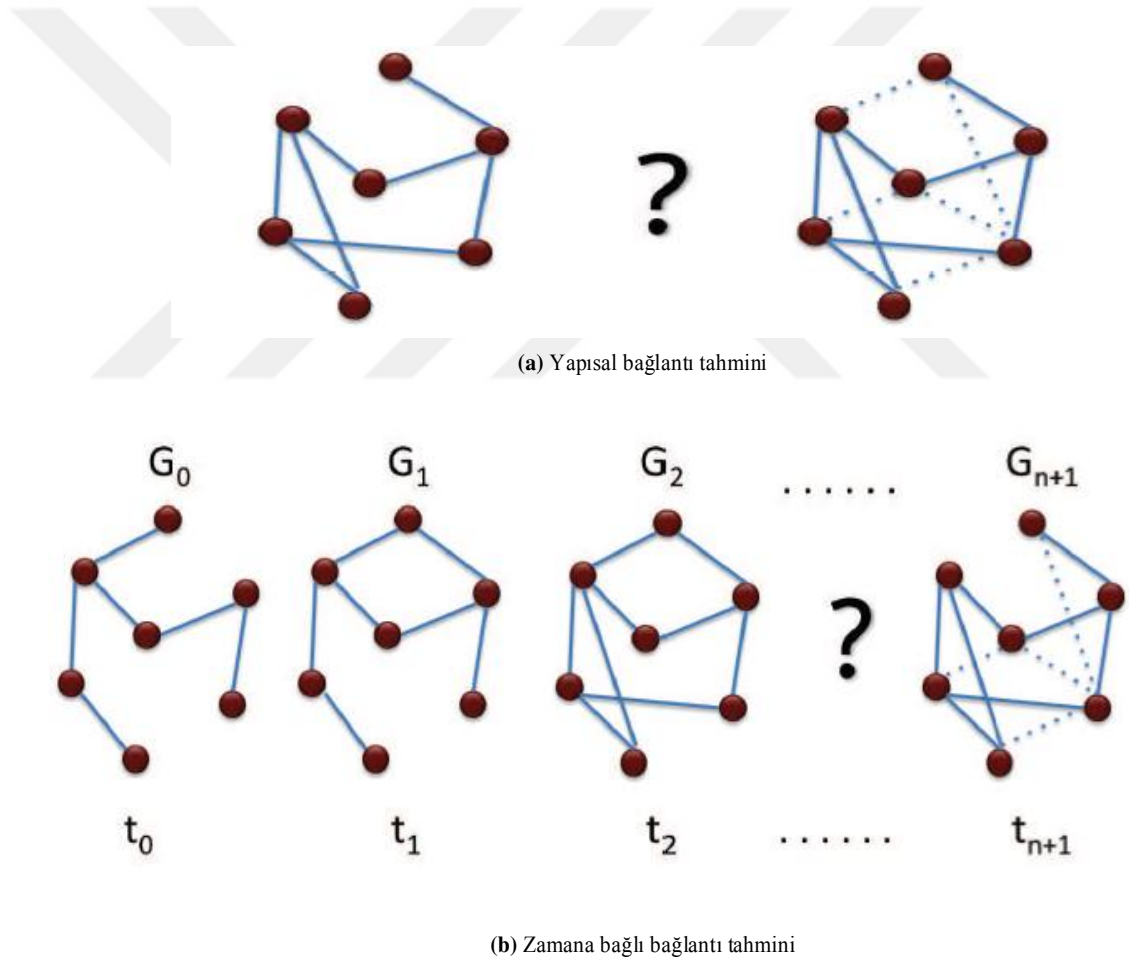
Sosyal ağlarda bağlantı tahmini problemi, modern bilgi biliminde uzun süredir var olan bir problemdir [72]. Bilgisayar bilimi topluluğu tarafından başlangıçta Markov zincirlerine [73,74] ve istatistiksel modellere [75] dayalı olasılıksal modeller önerildi. Sonraki zamanlarda daha çok ağ yapısına dayalı, alt graflara dayalı sıralama modeli [76], yerel bilgi modeli [77], gibi modeller kullanıldı. Ağ yapısına ve düğüm niteliklerine dayalı çalışmalardan ayrı ayrı iyi sonuçlar alındıkça bu iki yöntemi de kullanan yerel koşullu olasılık modelleriyle çalışmalar yapıldı [78].

Topolojik yaklaşımları kullanan bağlantı tahmini yaklaşımları, ağın evrimsel gelişiminin dikkate alınıp alınmaması gerçeğine dayanarak yapısal ve zamansal olarak sınıflandırılabilir. Şekil 3.2 iki tür bağlantı tahminini göstermektedir.

Yapısal bağlantı tahmini, mevcut bir ağda var olan eksik ya da gizli bağlantıların bulunması problemi ifade eder [7,79]. Temel olarak, ağdaki, mevcut verileri kullanarak doğrudan görülemeyen bağlantıları ortaya çıkarmayı hedefler. Kanser, HIV, Alzheimer vb. gibi çeşitli hastalıklar üzerindeki tıbbi çalışmalarda görülemeyen protein etkileşimlerini, gen örüntülerini bulmak için doğrudan uygulanabilir özelliğe sahiptir [80]. Ayrıca, sabıkalıların

sosyal medyadaki ya da telefon görüşmelerindeki ilişkileri incelenerek oluşabilecek suç eylemleri önceden sezilebilir [81].

Zamana bağlı bağlantı tahmini, bir ağın zamansal tarihini inceleyerek bir sonraki zaman diliminde oluşacak bağlantıları bulma problemini ifade eder [4,7,82]. Bu bağlantı tahmini yönteminde t zamanına kadar gözlemlenen ağ hakkında bilgi sahibiyiz ve amacımız $t + k$ gibi bir sonraki zamanda ortaya çıkabilecek yeni bağlantıları tahmin etmektir. Örnek olarak, e-ticaret alışveriş sitelerinde müşterilere bir sonraki alışverişlerinde alabilecekleri ürünler önerilebilir, Facebook, Twitter, Flickr gibi birçok sosyal medya sitesinde arkadaş tavsiyesi için kullanılabilir. Akademik bilgi ağlarında akademisyenlerin gelecekte hangi akademisyenlerle işbirliği kuracaklarını tahmin edilebilir [82,21].

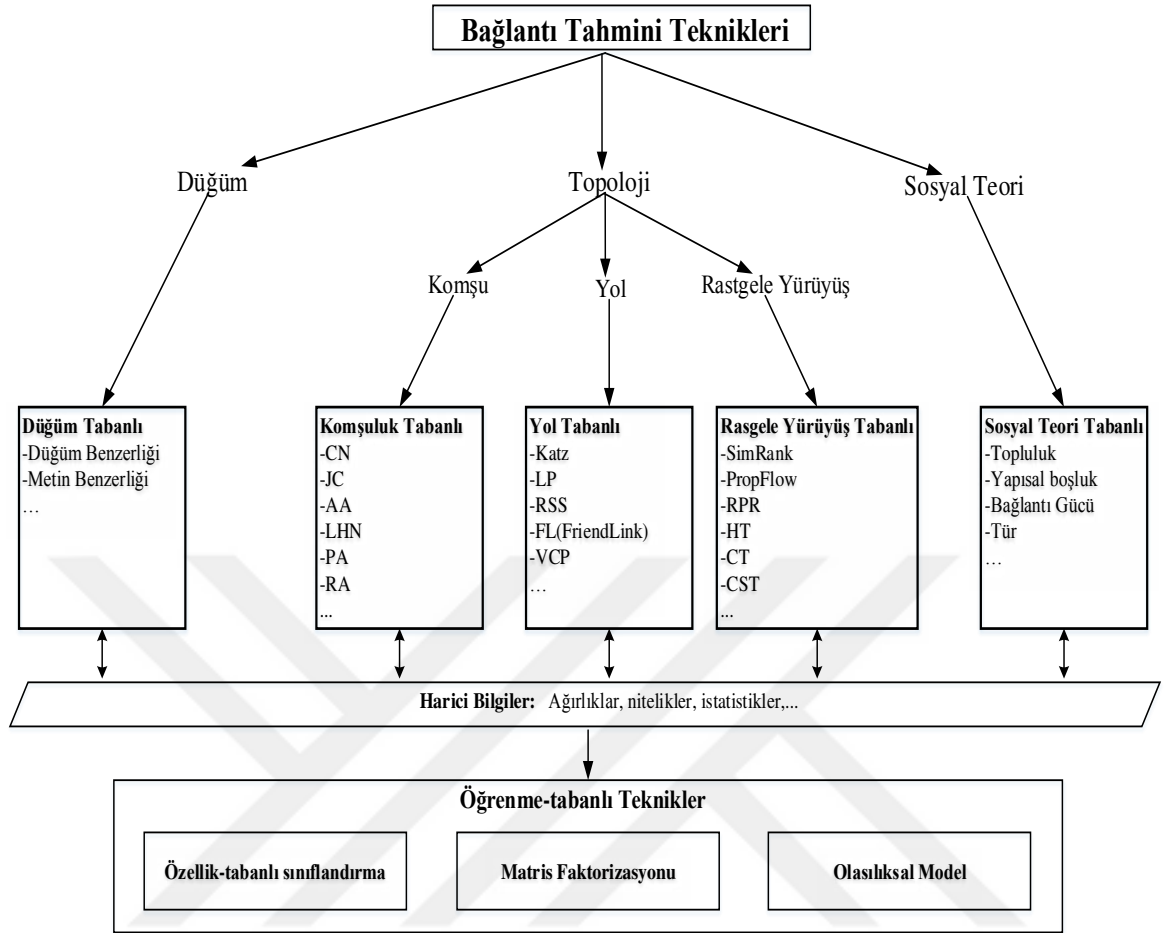


Şekil 3.2. Ağın topolojik özelliklerini kullanan yapısal ve zamansal bağlantı tahmini türleri [83]

Bağlantı tahmini probleminde, düğümler V (veri örnekleri) olarak düşünülürse $V = \{v_i\}_{i=1}^n, E$ ise bu veriler üzerinde var olan ilişkileri göstermektedir. Buna göre bir sosyal ağ

$G = (V, E)$ çizgesiyle tanımlanabilmektedir. Burada, (vi, vj) düğüm çiftleri olacaktır. Bu düğüm çiftleri arasında, $e_{ij} \notin E$ oluşmamış bağlantının tahmini yapılmaya çalışılmaktadır. Düğüm çiftleri arasındaki bağlantının önemini gösteren ölçüt, $skor(x, y)$ fonksiyonu olarak tanımlanabilmektedir. Literatürde sosyal ağlarda bağlantı tahmini için yaygın olarak kullanılan ölçütler, semantik veya topolojik tabanlı ölçütler olarak kategorize edilebilir. Semantik ölçütlerde düğümün içeriği benzerlik ölçütü olarak ele alınır. Örneğin yazar işbirliği ağında makalelerin anahtar kelimelerdeki benzerlikler kullanılarak yazarlar arasında gelecekteki etkileşimleri tahmin edilebilir [8]. Başka örneklerle açıklayacak olursak, örneğin ortak yazarlık ağında yazarların etnik kökenleri, eğitim bilgileri, akademik kariyerleri süresince çalıştıkları konu başlıkları arasındaki benzerlik kullanılarak bağlantı tahmini yapılabilir. Semantik ölçütlerden farklı olarak topolojik ölçütler benzerlik ölçütü hesaplamak için düğümlerin içeriklerinden ziyade ağın yapısını kullanırlar. İki düğüm arasındaki ortak komşuların sayısı topolojik ölçütlere bir örnektir. Semantik ölçütler için gerekli bilgiler çoğu zaman ağda mevcut olmayabilir veya analiz edilmesi masraflı olabilir. Topolojik ölçütler daha genel oldukları ve içerikle alakalı zengin bir özellik tanımına ihtiyaç duymadıkları için daha yaygın kullanılmaktadırlar [28]. Bazı yaklaşımlarda hem semantik hem de topolojik özellikler kullanılmıştır. Bu yaklaşımlar hibrit yaklaşımlar olarakta adlandırılabilir.

Ağdaki düğüm çiftlerinin benzerliklerini hesaplamak için düğüm, topolojik yapı ve sosyal teori bilgileri gibi topolojik özellikleri kullanan çok sayıda bağlantı tahmini ölçütü vardır. Ayrıca, bu üç kategoriden daha karmaşık olan öğrenme temelli bağlantı tahmini ölçütleri de bulunmaktadır. Öğrenme temelli ölçütlerde hem topolojik hem de semantik özellikler kullanılmaktadır. Bu bölümde, bağlantı tahmini ölçütleri ve yöntemleri için sistematik bir inceleme sunacağız. Bahsedilen sistematik yapı Şekil 3.3'te verilmiştir.



Şekil 3.3. Bağlantı tahmini teknikleri kategorileri

3.2.1. Düğüm Tabanlı Ölçütler

Bağlantı tahmini problemleri göz önüne alındığında, düşünülebilecek en sezgisel çözüm, verilen düğüm (aktör) çiftleri arasındaki benzerlik oranını hesaplamaktır. Aralarında bağlantı olmayan iki düğüm arasında bağlantı oluşma olasılığı, düğümler arasındaki benzerlikle doğru orantılıdır. Bu çalışmada mevcut ağda aralarında herhangi bir bağlantının olmadığı x ve y düğüm çifti arasındaki benzerliği gösteren nümerik değer $skor(x,y)$ ile temsil edilmiştir. Yüksek bir skor x ve y düğümlerinin gelecekte bağlanacağına dair yüksek olasılığı temsil ederken, düşük bir skor da bağlanmayacağına işaret etmektedir. Hesaplanan benzerlik skorlarının büyükten küçüğe sıralanmasını kullanarak, gelecekte oluşacak veya kaybolacak bağlantıları ya da mevcut ağda görünmeyen bağlantıları tahmin etmek mümkündür [84].

Düğüm ve düğümler arası ilişkilerden oluşan klasik bir sosyal ağda, düğümler çeşitli özelliklere sahip olabilir. Örneğin, sosyal medya uygulamalarında (Facebook, Twitter, Instagram vs.) kullanıcılar genellikle kariyer, adres, din ve ilgi alanları gibi özelliklere sahiptir. Bu özellikler, iki düğüm arasındaki benzerliği hesaplamak için doğrudan kullanılabilir. Çoğu durumda, düğümlerin öznitelik değerleri metin ya da dizge formatında tutulur. [85-87]'de yazarlar klasik metin tabanlı ve dizge tabanlı benzerlik ölçütleri hakkında detaylı bir araştırma sunmuşlardır.

Literatürde bu alanda yapılan birçok çalışma vardır. Anderson ve ark. [88] kullanıcılar arasında ki benzerliği hesaplamak amacıyla kullanıcıların ortak ilgi alanlarının kullanıldığı bir yöntem sundu. Çalışmada kullanıcıların ilgi alanlarını belirlemek için Wikipedia'da incelediği sayfaları ve Stack OverFlow'da sordukları soruları temel almışlardır. Çalışmada, kullanıcıların bu sitelerdeki her bir girişimi kullanıcı hareketi olarak kabul edilmiştir. Kullanıcı hareketleri vektörlerle temsil edilmiştir. Daha sonra kullanıcılar arasındaki benzerlik kullanıcıların ilgili eylem vektörleri arasındaki kosinüs benzerliği hesaplanarak bulunmuştur. Akcora ve ark. [89] sosyal medya kullanıcılarının çoğunun profil bilgilerinin eksikliğini fark etmiştir. Bunun için, kullanıcıların benzerliklerini ölçmeden önce bu eksik bilgilerin bir kısmının çıkarılması için bir yöntem önermişlerdir. Önerdikleri yöntem, arkadaşların ve anahtar kelimelerin sayısının artmasının, kullanıcı ve arkadaşları arasındaki ortalama benzerliği azalttığını göstermiştir. Sonuç olarak, düğüm tabanlı ölçütler temelde düğümlerin kişisel özelliklerini ve davranışlarını yansıtabilecek öznitelik ve eylemlere dayalıdır. Dolayısıyla, sadece bu gibi bilgilerin kolaylıkla edinildiği durumlarda, düğüm temelli ölçütler bağlantı tahmininde kullanışlı olmaktadır.

3.2.2. Topoloji Tabanlı Ölçütler

Daha öncede bahsettiğimiz gibi benzerlik ölçütleri semantik ve topolojik özellikler olarak ikiye ayrılır. Düğüm özelliklerinin kullanıldığı düğüm tabanlı ölçütlerde semantik özelliklerin bulunması zordur. Bu bölümde sunulan topolojik tabanlı ölçütler, düğüm ve kenar bilgisinden ziyade ağın topolojik (yapısal) bilgilerini kullanmaktadır. Son on yılda, düğüm ve kenarlar özniteliklerinin olmadığı basit bir ağda iki düğümün benzerliğini hesaplanması için topoloji tabanlı birçok ölçüt önerildi. Topolojik ölçütler literatürde genel olarak komşuluk tabanlı, yol tabanlı ve rastgele yürüyüş tabanlı olmak üzere üç başlık altında kategorize edilmiştir. Bu bölümde, bu kategorilerdeki en popüler bazı ölçütlerin genel bir

tanımı verilecektir. Tanımların daha iyi anlaşılması için bağlantı tahmini yöntemlerinde kullanılan bazı standart semboller Tablo 3.1’de verilmiştir.

Tablo 3.1. Bağlantı tahmini ölçütlerinde kullanılan ortak semboller

$\Gamma(x)$	x ’in komşular kümesi
$\Gamma(y)$	y ’in komşular kümesi
Büyük harfler	Sosyal ağın komşuluk matrisi
Küçük Harfler	Sosyal ağdaki düğümler
$ \Gamma(x) $	x düğümünün komşularının sayısı / derecesi
$ \Gamma(y) $	y düğümünün komşularının sayısı / derecesi

3.2.2.1. Komşuluk Tabanlı Ölçütler

Gerçek dünya verilerinden oluşturulmuş sosyal ağlarda, düğümler kendilerine benzeyen düğümlerle yeni bağlantılar kurma eğilimindedirler. Komşuluk tabanlı ölçütlerde, benzerlik oranı düğüm çiftinin ortak komşuları kullanılarak hesaplanır. Temel fikir x ve y düğümlerinin komşuları $\Gamma(x)$ ve $\Gamma(y)$ ne kadar büyük oranda kesişirse gelecekte aralarında bağlantı olma ihtimali de o kadar yüksektir. Bu alanda önerilmiş çok sayıda komşuluk tabanlı ölçüt bulunmaktadır.

Ortak Komşular (CN)

Ortak komşular ölçütü x ve y düğümleri için ortak komşuların sayısını göstermektedir. Basitliğinden dolayı bağlantı tahmini problemlerinde kullanılan en yaygın ölçüt haline gelmiştir [90]. Ortak komşular kullanılarak birçok farklı ölçüt türetilmiştir. Ortak komşuların sayısı ne kadar fazlaysa x ve y düğümleri arasında gelecek zamanda bağlantı oluşma olasılığı da o kadar yüksek olur. Bu ifadenin matematiksel karşılığı denklem (3.1)’de gösterildiği gibidir.

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)| \quad (3.1)$$

Ortak komşular ölçütü normalize edilmediğinden dolayı düğüm çiftleri arasındaki nispi benzerlikleri yansıtabilir. Bu ölçütten türetilmiş bazı ölçütler, ortak komşular ölçütünü normalleştirmeyi hedeflemiştir.

Jaccard Katsayısı (JK)

Jaccard katsayısı, ortak komşuların normalleştirilmiş halidir [72]. x ve y düğümlerinin tüm komşularının toplam sayısında ortak komşularının oranını hesaplar. Daha yüksek oranda pay alan düğüm çiftlerinin benzerliği daha fazladır. Bu ifadenin matematiksel karşılığı denklem (3.2)'de gösterildiği gibidir.

$$JK(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (3.2)$$

Adamic/Adar Katsayısı (AA)

Adamic/Adar katsayısı, Adamic ve Adar tarafından [91] ilk olarak iki web sayfasındaki içeriklerin birbirlerine ne kadar yakın olduğunu hesaplamak için önerilmiştir. Daha sonraları sosyal ağlarda da yaygın olarak kullanılmıştır. Bu ölçüt, ortak komşular (CN) ölçütünün aksine, daha az komşu sayısına sahip ortak komşuların ağırlığının daha yüksek olacağını savunur. Bu ifadenin matematiksel karşılığı denklem (3.3)'de gösterildiği gibidir.

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|} \quad (3.3)$$

Leicht-Holme-Nerman Katsayısı (LHN)

Leicht-Holme-Nerman katsayısı x ve y düğümlerinin ortak komşularının x ve y düğümlerinin derecelerinin çarpımına oranıdır [92]. Bu ifadenin matematiksel karşılığı denklem (3.4)'te gösterildiği gibidir.

$$LHN(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| \cdot |\Gamma(y)|} \quad (3.4)$$

Tercihli Bağlılık Katsayısı (PA)

Tercihli bağlılık ölçütü [93], yeni bağlantıların komşu sayısı fazla olan düğümler arasında oluşma ihtimalinin daha yüksek olduğunu savunur. Newman [90] ve Barabasi vd. [94]

çalışmalarına göre bir işbirliği ağında x ve y arasında işbirliği oluşma ihtimali x ve y 'nin iş birlikteliklerinin sayısı ile orantılıdır. Bu ifadenin matematiksel karşılığı denklem (3.5)'te gösterildiği gibidir.

$$PA(x, y) = |\Gamma(x)| * |\Gamma(y)| \quad (3.5)$$

Kaynak Paylaştırma Katsayısı (RA)

Bu ölçüt, Zhou ve ark. [77] tarafından karmaşık ağlarda kullanılmak üzere önerilmiştir. Kaynak paylaşımının fiziksel süreçlerinden esinlenmişlerdir. Önerilen yöntem, aralarında doğrudan bir etkileşim olmayan düğüm çiftleri arasındaki bağlantıları ölçer, diğer bir deyişle düğümlerin komşularını değil, komşularının komşularını düşünmektedir. Aralarında doğrudan bağlantı olmayan düğümler ortak komşuları sayesinde dolaylı olarak iletim sağlarlar. İletim sağlayan her düğümün kaynak ünitesi vardır ve bu kaynağı komşularına eşit dağıtır. Bu düğümler arasındaki benzerlik birbirlerinden aldıkları kaynaklara göre hesaplanır. Bu ifadenin matematiksel karşılığı denklem (3.6)'da gösterildiği gibidir.

$$RA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|} \quad (3.6)$$

RA ölçütü, AA ölçütüyle benzer özelliğe sahiptir. İkisi de yüksek dereceli komşuların katkısını azaltır. Ancak, RA ölçütü AA 'dan daha ağır bir şekilde cezalandırılır. Küçük çaplı ağlarda, AA ve RA ölçütleri birbirlerine yakın tahmin sonuçlarına sahiptir. Ancak, büyük çaplı ağlar için RA daha iyi bir performans göstermektedir.

Sørensen Katsayısı (Søren)

Bu ölçüt, ortak komşuların sayısını düşünmenin yanı sıra, daha düşük düğüm dercesine sahip düğüm noktalarının bağlantı kurma olasılıklarının daha yüksek olabileceğini savunur. Bu ifadenin matematiksel karşılığı denklem (3.7)'de gösterildiği gibidir [72].

$$Søren(x, y) = \frac{2|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| + |\Gamma(y)|} \quad (3.7)$$

Kosinüs Benzerliği Katsayısı (*Cos*)

Salton Kosinüs benzerliği olarakta adlandırılan bu ölçüt [72], x ve y düğümlerine sahip komşuluk matrisi satırları arasında kosinüs açısına dayanan benzerlik katsayısını bulmak için kullanılır. Bu ifadenin matematiksel karşılığı denklem (3.8)'de gösterildiği gibidir.

$$Cos(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)| \cdot |\Gamma(y)|}} \quad (3.8)$$

Hub Promoted Katsayısı (*HPI*)

Bu ölçüt x ve y düğümlerinin ortak komşularının x ve y düğümlerinin minimum derecelerine oranıyla hesaplanır [77]. Bu ifadenin matematiksel karşılığı denklem (3.9)'da gösterildiği gibidir.

$$HPI(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min\{|\Gamma(x)|, |\Gamma(y)|\}} \quad (3.9)$$

Hub Depressed Katsayısı (*HDI*)

Bu ölçüt x ve y düğümlerinin ortak komşularının x ve y düğümlerinin maksimum derecelerine oranıyla hesaplanır [77]. *HDI* ölçütü *HPI* ölçütüne göre daha düşük skor verir. Bu ifadenin matematiksel karşılığı denklem (3.10)'da gösterildiği gibidir.

$$HDI(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max\{|\Gamma(x)|, |\Gamma(y)|\}} \quad (3.10)$$

3.2.2.2. Yol Tabanlı Ölçütler

Yol tabanlı ölçütler, komşuluk tabanlı ölçütler gibi düğüm ve kenarların niteliklerini gerektirmeyen ölçüt türüdür. Bu ölçütler, iki düğüm arasındaki benzerliği hesaplamak için düğümler arasındaki yolların uzunluğunu kullanır [73].

Yerel Yol (*LP*)

LP ölçütü 2 ve 3 uzaklık mesafesindeki yerel yol bilgilerini kullanır [95]. En yakın komşuların bilgilerini kullanan benzerlik ölçütlerinden farklı olarak mevcut düğümden 3 mesafe uzaktaki komşularının ek bilgilerini kullanır. Şüphesiz, yol ne kadar kısa olursa düğümler arasındaki benzerlik o kadar yüksek olacağından dolayı uzaklık mesafesinin 2

olması 3 olmasından daha uygundur. Bu ölçütte daha tutarlı sonuçlar elde etmek için bir α ayarlama faktörü kullanılır. Bu faktör $-1 < \alpha < 1$ olmalıdır. Matematiksel denklemi (3.11)'de verilmiştir. Denklemden, A^2 ve A^3 düğümlere sırasıyla 2 ve 3 uzunluğundaki mesafelere sahip komşuluk matrisleridir.

$$LP = A^2 + \alpha \cdot A^3 \quad (3.11)$$

Katz Katsayısı (Katz)

Bu ölçüt, LP ölçütünden farklı olarak iki düğüm arasındaki tüm yolları dikkate alır. Leo Katz [10] tarafından önerilen bu ölçüt iki düğüm arasındaki en kısa yolların toplamını temel alır. Ölçütte, daha kısa yollara daha fazla ağırlık vermek amacıyla l yol uzunluğu önceden belirlenen sabit bir β parametresine üs olarak verilir. Bu ifadenin matematiksel karşılığı denklem (3.12)'de gösterildiği gibidir.

$$Katz(x, y) = \sum_{l=1}^{\infty} \beta^l \cdot |yollar_{x,y}^l| = \beta A + \beta^2 A^2 + \beta^3 A^3 + \dots \quad (3.12)$$

Denklemden $|yollar_{x,y}^l|$, x ve y düğümleri arasında olan l uzunluğundaki yolların sayısıdır ve $\beta > 0$ olmalıdır.

3.2.2.3. Rastgele Yürüyüş Tabanlı Ölçütler

Sosyal ağlarda düğümler arasında meydana gelen sosyal etkileşimler rastgele yürüyüşler olarak modellenebilir [96]. Sosyal ağlarda belirli bir düğümdeki rastgele bir yürüyüşçünün hedefi, bu düğümünden komşularına olan geçiş olasılıklarıyla gösterilir. Literatürde, düğümler arasındaki benzerlikleri rastgele yürüyüşe dayalı hesaplayan birçok bağlantı tahmini ölçütü vardır.

Ulaşma Zamanı (HT)

Ulaşma zamanı ölçütünün temel fikri çizge teorisinden gelir. Burada, x başlangıç düğümünden hedef y düğüme erişmek için yapılan rastgele bir yürüyüşteki adım sayısı hesaplanır. Ulaşma zamanının düşük olması x ve y düğümlerinin benzer olduğunu

gelecekte aralarında bağlantı oluşma ihtimalinin yüksek olduğunu gösterir [96]. Bu ifadenin matematiksel karşılığı denklem (3.13)'te gösterildiği gibidir.

$$HT(x, y) = 1 + \sum_{\omega \in \Gamma(x)} P_{x,\omega} HT(\omega, y) \quad (3.13)$$

$P = D_A^{-1}$, A 'nın D_A köşegen matrisinde $(D_A)_{i,i} = \sum_j A_{i,j}$ değerleriyle oluşturulur ve $P_{i,j}$, i düğümünden j düğümüne ulaşana kadar atılacak adımların olasılığıdır.

Gidip Gelme Zamanı (CT)

HT ölçütü simetrik değildir. Bu sebeple, CT ölçütü, x düğümünden y 'ye ve y düğümünden x 'e ulaşmak için atılması beklenen adımları sayar. Bu ifadenin matematiksel karşılığı denklem (3.14)'te gösterildiği gibidir.

$$CT(x, y) = HT(x, y) + HT(y, x) \quad (3.14)$$

Köklü PageRank (RPR)

Bu ölçüt [7], arama motorları tarafından arama sonuçlarının sıralanması için kullanılan PageRank algoritmasından türetilmiştir. Çizgede bir düğümün sırası, bu düğüme çizgedeki rastgele yürüyüşle erişebilme olasılığıyla orantılıdır. Ayrıca x ve y arasındaki rastgele yürüyüş her bir adımda α parametresindeki olasılık değeri ile yeniden başa döndürülebilir. Böylece ağı çok uzak kısımlarının gezilmesinden kurtulmuş olunur. Rastgele yürüyüş β $[0,1]$ olasılığı ile başa döner, $1 - \beta$ olasılığı ile o an bulunan düğümün komşularından rastgele birine gider. Bu işlem her adımda uygulanır. D matrisi $D_{i,i} = \sum_j A_{i,j}$ ile oluşturulmuş köşegen matristir. $N = D^{-1}A$, komşuluk matrisinin satırlarının bir normalleştirilmesidir. Bu ifadenin matematiksel karşılığı denklem (3.15)'te gösterildiği gibidir.

$$RPR = (1 - \beta)(I - \beta N)^{-1} \quad (3.15)$$

3.2.3. Sosyal Teori Tabanlı Ölçütler

Son yıllarda, sosyal ağ madenciliği ve analizi problemlerini çözmek amacıyla çok sayıda çalışma yapılmaktadır. Gün geçtikçe bu alanda yapılan çalışmaların çoğu topluluk, zayıf ve güçlü bağlantı kavramları, homojen ve yapısal denge gibi klasik sosyal teorileri kullanıyor. Yalnızca düğüm ve topolojik özellikleri kullanan önceki ölçütlerin aksine, sosyal teoriye dayanan bağlantı tahmini ölçütleri özellikle büyük ölçekli sosyal ağlar için yararlı sosyal etkileşim bilgilerini açığa çıkararak tahmin performansını arttırabilir.

Mark ve ark. [97], insanlar arasındaki bağlantıları etkileyen birçok faktör olduğunu öne sürmüşlerdir. Çalışmalarında, insanlar arası ilişkileri zayıf ve güçlü olmak üzere iki gruba ayırmışlardır. Güçlü bağlar akrabalık, arkadaşlık vb. ilişkileri ifade ederken, zayıf bağlar ise toplumdaki diğer ilişkileri işaret eder. Zayıf bağların, büyük ölçekli ağlarda faydalı olabileceğini savunmuşlardır.

Valverde-Rebaza ve Lopes [98], Twitter kullanıcılarının hem topolojik bilgilerini hem de ilgi ve davranışlarını göz önünde bulundurarak gelecekte kuracakları olası bağlantıları tahmin etmeye çalışmışlardır. Çalışmanın deneysel sonuçları, önerdikleri yöntemin büyük ölçekli, yönlü ve asimetrik sosyal ağlardaki tahmin performansını etkili bir şekilde etkilediği kanıtlanmıştır.

Qiu ve ark. [99], sosyal ağlardaki büyüme dinamiklerini yakalayabilmek için ağın evrimsel gelişimine dayalı bir model önermişlerdir. Bu model, gerçek ağlarda gözlemlenen önemli ağ yapılarını ortaya çıkardığından ağın büyüme sürecinin daha iyi karakterize edilebileceğini savunur. Daha sonra ise bağlantı tahmini doğruluğunu geliştirmek için düğüm davranışları kullanılır.

Önerilen bu modeller karmaşık bağlantı tahmin yöntemlerini tasarlamak için kullanılabilir.

3.2.4. Öğrenme Tabanlı Ölçütler

Öğrenme tabanlı bağlantı tahmini ölçütleri, hem düğüm tabanlı hem de topoloji tabanlı özelliklere dayanır. Son yıllarda düğümün özelliklerini ve harici bilgilerini kullanan birçok öğrenme temelli bağlantı tahmini yöntemi önerilmiştir [26,86]. Bu öğrenme temelli yöntemler, özellik tabanlı sınıflandırma, olasılıklı grafik modeli ve matris faktörizasyonu olarak kategorilere ayrılabilir.

Öğrenme temelli yöntemlerin çoğu tipik bir özellik sınıflaması problemi olarak kabul edilebilir. Ichise ve ark. [100], araştırmacıların akademik çalışmalarını kullanarak oluşturduğu ortak yazarlık ağında bağlantı tahminini iyileştirmek için soyut bilgiler, çalışma başlıkları, yayınlanma tarihi gibi bilgileri kullanarak öğrenme tabanlı semantik bir yaklaşım ortaya koymuşlardır. Scellato ve ark. [6], konum tabanlı bir sosyal ağda bağlantı tahmini kalitesini arttırmak için mekan bilgisi ve küresel özellikler gibi sosyal özellikleri kullanmışlardır. Scripps ve ark. [101], bağlantı tahmini için matris sıralamasına dayalı ayırt edici bir öğrenme tekniği sunmuşlardır. Önerilen bu yöntem sayesinde, bir ağın komşuluk matrisini düğüm öznitelikleri ve komşularının topolojik özellikleriyle hesaplanan ağırlıklı matrislere dönüştürerek belirgin öznitelikler ve topolojik özellikler otomatik olarak belirlenebilir.

Sosyal ağlarda, her bir düğüm çifti arasındaki bağlantıya topolojik benzerlik ya da rastgele yürüyüşte geçiş olasılığı gibi bir olasılık değeri atanabilir. Bu yöntemle oluşturulmuş çizgelere olasılıksal çizge modeli denir. Literatürde olasılıksal çizge modeline dayalı öğrenme tabanlı birçok bağlantı tahmini yöntemi vardır. Clauset ve ark. [102] sosyal ağdaki hiyerarşik yapıyı çıkaracak bir model önerdiler ve önerdikleri yöntemi bağlantı tahmini problemi için uyguladılar. Kashima ve ark. [103], iki düğüm arasında bağlantı olup olmayacağını tahmin etmek için ağın evrimsel gelişiminin olasılıksal parametrelerle ifade edildiği bir model önerdiler.

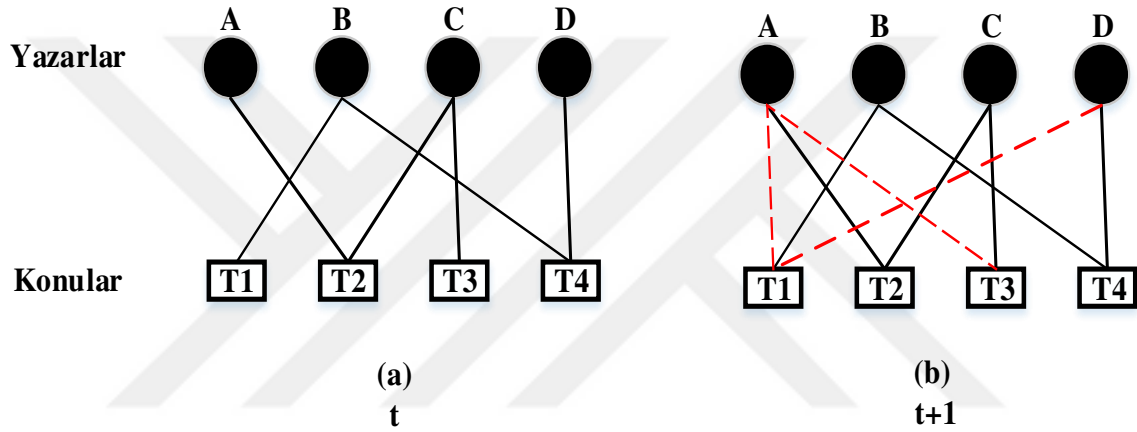
Bağlantı tahmini problemi, matris tamamlama diğer bir adıyla matris faktörizasyonu olarak da görülebilir. Menon ve ark. [79], bağlantı tahmini problemini matris tamamlama problemi olarak ele almış ve bağlantı tahmini problemini çözmek için matris faktörizasyon yöntemini geliştirmişlerdir.

3.3. İki Parçalı Ağlarda Bağlantı Tahmini

Günümüzde uygulama alanları giderek artış gösteren sosyal ağlardaki ilerlemeyle “Ağ” ailesine birçok yeni üye eklenmiştir. Bunlardan biri de farklı işlevlere sahip iki farklı düğüm kümesi içeren iki parçalı ağlardır [19,20]. İki parçalı ağlarda bağlantı tahmini mevcut ağda olmayan gelecekte oluşacak bağlantıların bulunmasına eşittir. $G = (U, V, E)$ üçlü çizge, t zamanına kadar gözlemlenen iki parçalı bir ağı temsil eder. Bağlantı tahmini doğrudan t zamanında ağda bulunmayan ancak $t + 1$ zamanında mevcut bağlantıları bulmayı amaçlar. Matematiksel olarak tanımlarsak, bağlantı tahmini görevi, t zamanında $u \in U$, $v \in V$

ve $(u, v) \notin E$ olduğu G iki parçalı ağda, $t + 1$ zamanında u ve v düğümleri arasında (u, v) bağlantısının oluşup oluşmayacağını tahmin etmektir [104].

Şekil 3.4, ayrı düğüm kümelerinden birinin yazarlar ve diğerinin ise bu yazarların çalıştıkları konu etiketleri olduğu iki parçalı bir ağda bağlantı tahmini problemini gösterir. Kesikli çizgi bir sonraki zaman diliminde yazarın çalışması öngörülen konu etiketleriyle olan bağlantıları, kesintisiz çizgiler ise yazarın t zamanına kadar çalıştığı konu etiketleriyle olan bağlantıları göstermektedir. Bağlantı tahmininin asıl amacı, ağdaki bağlantı bilgilerine (kesintisiz çizgiler) göre yazar-konu ağında olası bağlantıları (kesikli çizgiler) tahmin etmektir.



Şekil 3.4. İki parçalı ağlarda bağlantı tahmini örneği

İki parçalı ağlarda bağlantı tahmini probleminde ilk ve doğal girişim, iki parçalı ağın tek parçalı ağa yansıtılması ve daha sonra geleneksel tek parçalı ağlarda kullanılan yöntemler ile analiz edilmesidir. Yansıtılmış diğer bir deyişle izdüşümü alınmış bir ağ, bilinen klasik tek parçalı ağ yapısıdır. Yansıtılmış bir ağda bağlantı tahmininde bulunmak için, ağın doğrudan topolojik özelliklerini kullanan bir model geliştirmek mümkündür ve ağ analizleri için sıkça yapılan bir işlemdir.

Bu çalışmada potansiyel bağlantıları analiz etmek amacıyla ilk olarak iki parçalı ağ, yansıtılmış ağ olarak adlandırılan tek parçalı ağlara dönüştürülür. Tek parçalı ağlara dönüştürme teknikleri Bölüm 2.4.2’de detaylı olarak anlatılmıştır.

3.3.1. İki Parçalı Ağlarda Bağlantı Tahmini Yöntemleri

Bu bölümde, literatürde iki parçalı ağlarda bağlantı tahmini problemi için önerilmiş yaygın olarak kullanılan iki temel yöntem sunulacaktır: İç bağlantılar yöntemi ve işbirlikçi filtreleme yöntemi.

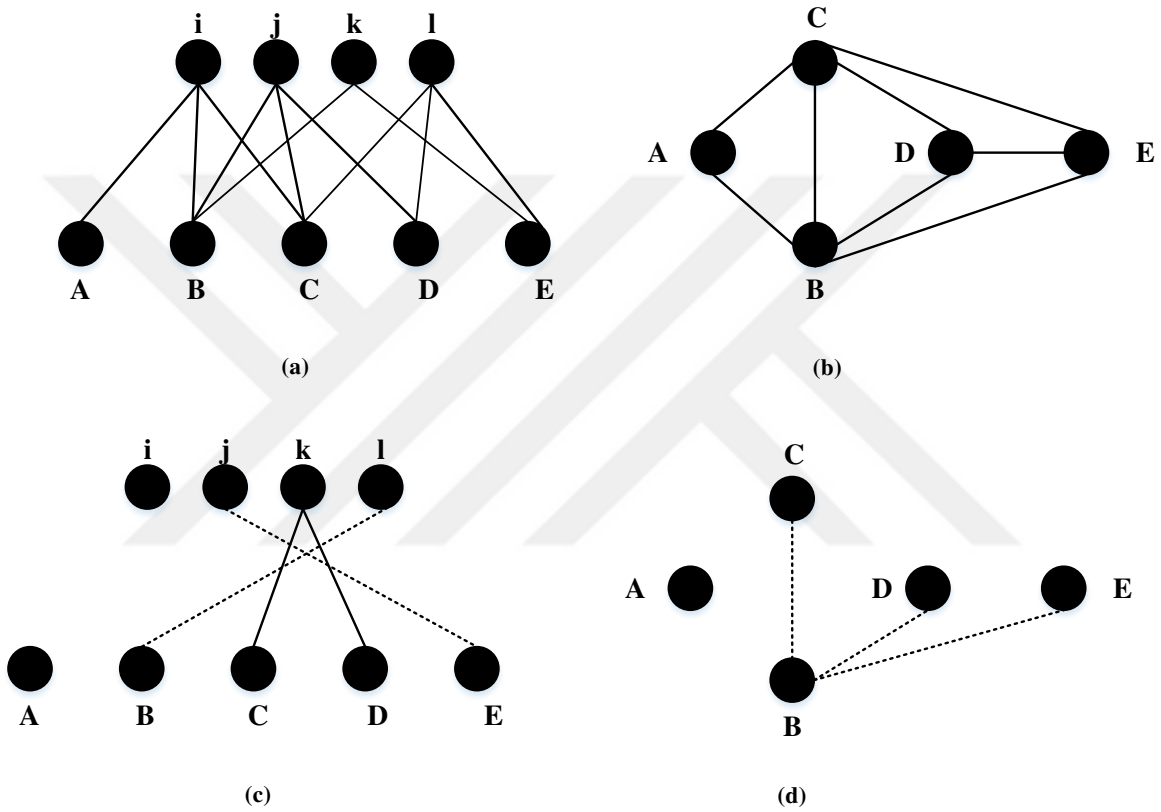
3.3.1.1. İç Bağlantılar Tahmini Yöntemi

Gerçek dünya ağlarının doğal yapısını ifade eden iki parçalı ağlarda bağlantı tahmini için kullanılan yöntemlerden en önemlisi iç bağlantı tahmini olarak adlandırılan yöntemdir. Bu bağlantı tahmini yönteminin en temel özelliği, iç bağlantılara odaklandığı için bağlantı tahminin de ağda olmayan tüm bağlantıları ele almak yerine sadece iç bağlantı olma özelliğine sahip bağlantıları dikkate almasıdır. Bu sayede öngörülen bağlantı sayısı azalmakta, buna paralel olarak tahminlerinin kalitesi de artmaktadır. İki parçalı ağı oluşturan düğümler alt ve üst düğümler olarak düşünülürse, G iki parçalı ağda var olan iki alt düğüm arasında hâlihazırda bir etkileşim varsa muhtemelen gelecekte de bu iki düğüm arasında etkileşim olacaktır. Öte yandan, G 'de ortak bir komşusu olmayan iki alt düğümün gelecekte etkileşim içerisine girmesi düşük bir ihtimaldir. Bu durum için bir ağırlık fonksiyonu kullanılarak daha somut analizler yapılabilir. Öngörülen bağlantıların, yüksek ağırlıklarla bağlantılı "iç bağlantılar" olduğunu varsayarsak bu tahminlerin oranları daha yüksek olacaktır.

Tanım 4 (İç Bağlantı Özelliği): $G = (U, V, E)$ iki parçalı ağda $u \in U$ ve $v \in V$ düğümleri arasında bağlantı olmayan $(u, v) \notin E$ iki düğümdür. $G' = (U, V, E)$, $E' = E \cup \{(u, v)\}$ ise iki parçalı G ağına, $(u, v) \notin E$, $(u, v) \in U \times V$ bağlantısının eklenmesiyle elde edilen iki parçalı ağdır. Bu düğüm çiftinin G ağına eklenmesiyle oluşan G' ağının yansıtılmış tek parçalı ağında G'_U herhangi bir değişiklik yaratmaması gerekir, diğer bir deyişle $G'_U = G_U$ olmalıdır. Sözel olarak ifade etmek gerekirse, (u, v) düğüm çiftinin G iki parçalı ağda iç bağlantı özelliğine sahip olması için bu bağlantının eklenmesi U - yansıtılmış ağda hiçbir değişikliğe sebep olmaması gerekir.

Şekil 3.5'te iç bağlantı tahmini örneği gösterilmiştir. G iki parçalı ağ yapısal olarak incelendiğinde elde edilen sonuçlara göre, iç bağlantı kümesi:

$\{(B, l), (C, k), (D, k), (E, j)\}$ 'dir. (B, l) düğüm çiftini ele alalım: l düğümünün iki parçalı ağdaki komşuları $\Gamma(l) = (C, D, E)$ düğümleridir. G iki parçalı ağında $\Gamma(B) = (i, j, k)$ komşularına sahip B düğümü, Şekil 3.5(b)'de görüldüğü gibi $(B, C), (B, D), (B, A)$ ve (B, E) düğümleriyle bağlantılıdır. Bu sebeple (B, l) bağlantı tahmininin ağa eklenmesi Şekil 3.5(b) yansıtılmış ağda herhangi bir değişiklik yaratmayıp sadece ortak komşu sayısını 2 ye çıkaracağından, (B, l) bağlantısının iç bağlantı özelliğine sahip olduğunu söyleyebiliriz.



Şekil 3.5: İç bağlantı tahmini örneği. (a) $G = (U, V, E)$ iki parçalı bir ağ, (b) G iki parçalı ağının U -yansıtılmış $G_U = (U, E_U)$ tek parçalı ağı, (c) G iki parçalı ağdaki iç bağlantılar, (d) (B, l) iç bağlantısı tarafından oluşturulmuş U -yansıtılmış ağdaki bağlantılar [43].

3.3.1.2. İşbirlikçi Filtreleme Yöntemi

Literatürde ilk olarak 1992 yılında Goldberg [104] tarafından önerilen İşbirlikçi filtreleme yöntemleri daha çok iki parçalı ağlarda tavsiye edici sistem uygulamalarında kullanılır. Sarwar ve arkadaşları [105], işbirlikçi bir filtreleme algoritmasıyla aktif kullanıcı ile benzer olan kullanıcıların söz konusu ürüne verdikleri değerlendirmeleri kullanarak, aktif kullanıcının söz konusu ürüne verdiği değeri hesaplamaya çalışmışlardır. İki parçalı ağlarda,

klasik işbirlikçi filtreleme yaklaşımları U düğümlerinin kendilerine benzeyen U -düğüm kümesindeki komşularıyla bağlantı oluşturma eğilimlerini tahmin etmeyi hedefler. Örneğin, iki parçalı bir müşteri-ürün alışveriş ağında müşteriler büyük ihtimalle kendilerine benzeyen müşterilerin aldıkları ürünleri alacaklardır [66]. Daha açıklayıcı olursak, işbirlikçi filtreleme yöntemleri ilk olarak iki parçalı ağın U - düğüm kümesindeki her bir düğümün U -yansıtılmış ağda kendisine en çok benzeyen k adet komşularının iki parçalı ağdaki bu düğümlere güçlü bağlarla bağlanmış N adet komşularını seçer. Burada, U kümesindeki düğümlerin kendilerine benzeyen komşuları önceden oluşturulmuş U -yansıtılmış ağırlıklı ağlar kullanılarak bulunur [66]. Özetlemek gerekirse, işbirlikçi filtreleme yöntemi, her N düğümü için en yüksek puanlara sahip bağlantıları gelecekte oluşması muhtemel potansiyel bağlantılar olarak kabul eder.

İki parçalı ağda genellikle [43]'de önerilen işbirlikçi filtreleme yöntemi uygulanır: $G(U, V, E)$ iki parçalı bir ağ, $G_U = (U, E_U)$ iki parçalı ağdan çıkartılmış U -yansıtılmış ağ ve w ise ağırlık fonksiyonu olsun. Burada, her bir $u \in U$ düğümü için, $U_k \subseteq \Gamma_U(u)$, U -yansıtılmış tek parçalı ağdaki en yüksek kenar ağırlığına sahip k adet komşu seti dikkate alınır. Daha sonra, her bir $v \in \Gamma(U_k) \setminus \Gamma(u)$ için (u, v) bağlantı skoru hesaplanır. $skor(u, v)$, matematiksel hesabı aşağıdaki gibidir:

$$skor(u, v) = \sum_{z \in U_k \cap \Gamma(u)} \omega(u, z). \quad (3.16)$$

3.3.2. İki Parçalı Ağlarda Bağlantı Tahmini Problemini Ele Alan Çalışmalar

Allai ve ark. [106]'na göre müşteri-ürün gibi çoğu gerçek dünya karmaşık veri ağları doğası gereği zamanla gelişen iki parçalı yapıya sahiptir. Bağlantı tahmini bu ağların yapısını anlamak için popülerleşmiş ana yaklaşımlardan biridir. Pratikte iyi sonuçlar alınmasına rağmen sadece birkaç çalışma iki parçalı ağları ele almıştır. Bu çalışmada iki parçalı ağlarda adlandırılan iç bağlantılar kavramı tanımlanmıştır. Ve bu kavrama dayalı bir bağlantı tahmini yöntemi önerilmiştir. Algoritma, DBLP yazar-konferans veri kümesi, Flickr ve Delicious veri kümeleri üzerinde test edilmiştir. Önerdikleri yöntemde mevcut iki parçalı ağda bulunmayan tüm bağlantıları ele alıp daha sonra önceden belirlenmiş eşik değerinden büyük iç bağlantı ağırlığına sahip olanları potansiyel bağlantılar olarak belirler. Fakat bu çalışmanın en büyük dezavantajı uygun eşik değerini seçmektir.

Goa ve ark. [11], Allai ve arkadaşlarının kullandığı yöntemden farklı olarak mevcut ağda bulunmayan tüm bağlantıları potansiyel bağlantı olarak ele almak yerine mevcut ağdan aday anahtar çiftleri çıkaran daha sonra bu çiftler arasında olmayan eksik bağlantıları potansiyel bağlantı statüsüne alan yeni bir yöntem önerdiler.

Shibata ve ark. [107], farklı özelliklerde sosyal ağlar için birden çok öğrenme yaklaşımı ve bağlantı tahmini modeli kullanarak atıf tahmininde bulunan bir yöntem önermişlerdir.

Benchettara ve ark. [17], iki parçalı büyük çaplı sosyal ağlarda bağlantı tahmini problemini ele almışlardır. Bu problemi ele alırken, bağlantı tahmini probleminin iki farklı varyasyonu üzerinde çalışılmıştır: İki parçalı ağlarda bağlantı tahmini ve iki parçalı ağı oluşturan düğüm kümelerinden elde edilen tek parçalı ağlarda bağlantı tahmini. Bu çalışmada önerilen yöntem iki gerçek veri kümesi üzerinde test edilmiştir. Bunlardan biri DBLP bibliyografik veri tabanı kullanarak elde edilen ortak yazarlık ağı, diğeri ise çevrimiçi e-ticaret müzik sitesinin sekiz yıllık işlem geçmişinden elde edilen ağıdır.

Lie ve ark. [19], çekirdek tabanlı makina öğrenmesi yaklaşımıyla iki parçalı ağlarda kullanılan tavsiye edici sistem önermişlerdir.

Yiwen Cui ve ark. [108], heterojen ağlarda her bir bağlantı için hesaplanmış kenar ağırlığını kullanan bağlantı tahmin yöntemi önermişlerdir.

Elham ve ark. [109], bir bireyin herhangi bir çalışma alanına ilgisi olup olmadığını bulmaya çalışmışlardır. DBLP veri tabanından aldıkları bibliyografik verileri kullanarak iki parçalı ağ oluşturmuşlardır. Diğer çalışmalardan farklı olarak ağın ayrık kümelerini Spectral Graph Clustering algoritmasını kullanarak kendi içinde kümelemiş, daha sonra bu kümeleneş iki parçalı ağa Bipartite Graph (Bigraph) Spectral Co-clustering algoritmasını uygulamışlardır. Kümeleme ve Co-Clustering yöntemlerinin avantajlarını eşzamanlı kullanarak yüksek doğruluk oranlarında tahminler elde etmişlerdir.

Shuang ve ark. [23], çalışmalarında geleneksel sosyal ağ madenciliği yöntemlerinin yanında (Ortak Komşuluk, Jaccard, Adamic Adar vb.), yapısal boşluklar kavramını da hesaba katmıştır. Daha önceden hiç hesaba katılmayan bu kavram sayesinde yüksek doğruluk değerine sahip bağlantı tahminleri elde edilmiştir. Deneysel sonuçlar IMDb veri tabanından alınan verilerle oluşturulan iki parçalı ağ üzerinde test edilmiştir. Önerilen yöntem diğer metotlarla karşılaştırıldığında en iyi sonuçlara sahip olduğu kanıtlanmıştır.

Tsunghan ve ark. [110], diğer benzer çalışmaların aksine ağın doğasındaki dinamik yapıyı kullanmak yerine zamansal bilgileri temel alarak iki parçalı ağlar üzerinde bağlantı tahmini yöntemi önermişlerdir.

3.4. Bağlantı Tahmini Değerlendirme Ölçütleri

Klasik olarak bağlantı tahmini algoritmalarının performanslarının değerlendirilmesi için önerilmiş çeşitli değerlendirme ölçütleri bulunmaktadır. Bu ölçütlerden en yaygın kullanılanlar aşağıda sıralanmıştır. Değerlendirme ölçütlerinin daha iyi anlaşılması için bazı standart semboller Tablo 3.2’de verilmiştir.

Tablo 3.2. Bağlantı tahmini değerlendirme ölçütlerinde kullanılan terimler

<i>DP</i> (Doğru Pozitif)	Bağlantı tahmini algoritmasının pozitif olarak etiketlediği ve test ağında var olan bağlantıların sayısıdır.
<i>YN</i> (Yanlış Negatif)	Bağlantı tahmini algoritmasının negatif olarak etiketlediği ve test ağında olmayan bağlantıların sayısıdır.
<i>YP</i> (Yanlış Pozitif)	Bağlantı tahmini algoritmasının negatif olarak etiketlediği ve test ağında var olan bağlantıların sayısıdır.
<i>DN</i> (Doğru Negatif)	Bağlantı tahmini algoritmasının pozitif olarak etiketlediği ve test ağında olmayan bağlantıların sayısıdır.

Kesinlik: Bir bağlantı tahmini algoritmasının kalitesini değerlendirmek için, doğru tahmin edilen *DP* pozitif gözlemlerin yapılan toplam tahmin sayısına oranıdır. Kesinlik puanı ne kadar yüksekse, tahmin algoritmasının o kadar etkili olduğu anlamına gelir.

$$Kesinlik = \frac{DP}{DP+YP} \quad (3.17)$$

Alternatif olarak, hesaplanan bağlantı tahmini puanlarına göre sıralanan *top – n* bağlantı için Kesinlik ölçütü aşağıdaki gibi hesaplanır.

$$Kesinlik_n = \frac{m}{n} \quad (3.18)$$

m parametresinin değeri *top – n* bağlantı içerisinde ki *DP* sayısıdır.

Anma: Anma, doğru tahmin edilen pozitif gözlemlerin, gerçekte olması gereken tüm bağlantıların toplam sayısına oranıdır. Duyarlılık olarakta adlandırılabilir.

$$Anma = \frac{DP}{DP+YN} \quad (3.19)$$

F1-Değerlendirme Ölçütü: Kesinlik ve Anma değerlerinin harmonik ortalamasıdır.

$$F = \frac{2*Kesinlik*Anma}{Kesinlik+Anma} \quad (3.20)$$

Yukarıda bahsedilen bu değerlendirme ölçütleri performansı hesaplamak için sabit bir eşik (k) değeri kullanır. Bu sınırlı ölçütler doğru seçilmediği zaman yüksek veya düşük performans değerleri gösterebilir. Bu gibi durumlarla başa çıkmak için eşik eğrisi tabanlı ölçütler kullanılabilir. Bu değerlendirme ölçütleri aşağıda verilmiştir.

ROC Eğrileri: Farklı eşik değerlerinde doğru pozitif oranı (DPO) ile yanlış pozitif oranı (YPO) arasındaki eğilimi gösterir.

$$DPO = \frac{DP}{DP+YN} \quad (3.21)$$

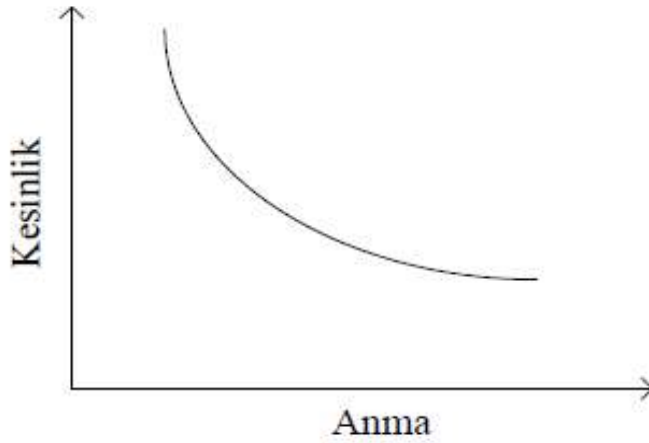
$$YPO = \frac{DP}{DN+YP} \quad (3.22)$$

AUC Eğrisi (Eğri Altında Kalan Alan): Bağlantı tahmini algoritmaları ile mevcut ağda olmayan fakat gelecekte oluşması öngörülen tüm bağlantıların benzerliklerini hesapladıktan sonra, AUC değeri rastgele seçilen mevcut bir bağlantıya verilen puanın, rastgele seçilmiş var olmayan bir bağlantıya verilen puandan daha yüksek olma olasılığı olarak yorumlanabilir. Her seferinde rastgele bir tane mevcut bağlantı ve bir tane de var olmayan bağlantı seçilerek benzerlik skorları hesaplanır. n tane karşılaştırma için n' var olan bağlantının benzerlik puanının var olmayan bağlantının benzerlik puanından daha büyük olduğu karşılaştırma sayısı, n'' ise eşit veya küçük olduğu karşılaştırma sayısıdır. Buna göre AUC değeri aşağıdaki gibi tanımlanır:

$$AUC = \frac{n' + 0.5n''}{n} \quad (3.23)$$

Genellikle, daha büyük bir AUC değeri, tahmini sonucunun daha yüksek kalitede olduğunu gösterir. Açıkçası, bir bağlantı tahmini sonucu için AUC değeri $AUC \in [0, 1]$ 'i karşılar. Bu nedenle, en iyi tahmini sonucun AUC değeri 1'dir ve sonucun rastgele tahmini ile AUC değeri 0.5'dir. Bu nedenle, 0.5'in altında bir AUC değerine sahip bir tahmin sonucu geçersizdir.

Kesinlik-Anma Eğrileri (PR Eğrisi): PR eğrilerinde her bir nokta farklı n eşik değerindeki kesinlik ve anma değerine denk gelir. x -ekseni anma ve y -ekseni hassasiyettir. Kesinlik ve anma arasında ters orantı vardır.



Şekil 3.6. Kesinlik ve anma arasındaki ilişki [22]

ROC ve PR değerlendirme ölçütleri arasındaki en önemli fark bir PR eğrisinin gerçek negatifleri hesaba katmamasıdır. Bu sebeple PR eğrileri, özellikle gerçek pozitif örneklerin tahmininde algoritmaların performansları arasındaki farkı göstermede çok daha iyidir. ROC ve AUC değerlendirme ölçütleri algoritmalar arasındaki performans değerlendirmelerinde çok küçük bir fark gösterecektir. Bir modelin pozitif ve negatif sınıflarda eşit derecede iyi performans göstermesi gerekiyorsa, ROC daha çok tercih edilir [111].

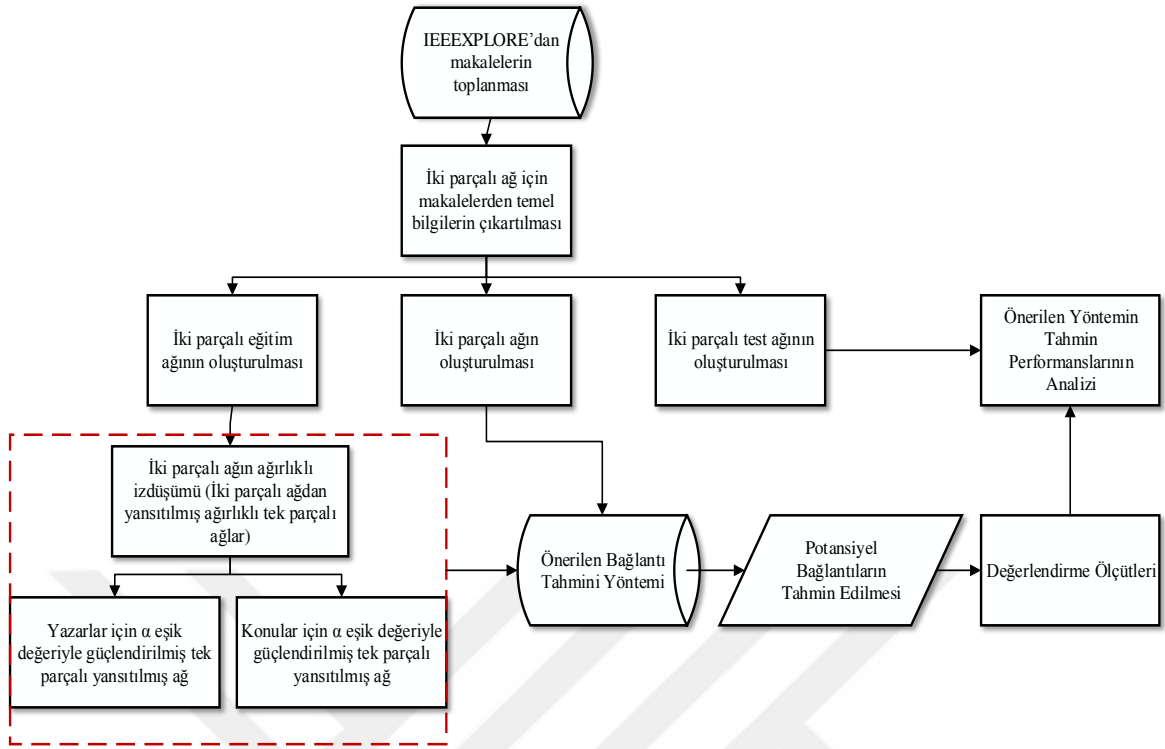
Birbirinden farklı bağlantı tahmini yöntemlerinin kalitesini değerlendirmek amacıyla yaygın olarak kullanılan en temel iki ölçüt: AUC (Eğri Altında Kalan Alan) [72] ve kesinlik [112] ölçütleridir. Hem kesinlik hem de AUC değerlendirme ölçütleri, algoritmaların tahmin kalitesini değerlendirirken sadece $top - n$ sonuçlarının kalitesini göz önüne aldığından

deneysel sonuçlar sınırlı olabilir. Bu sebeple, bu tez çalışmasında algoritmaların bağlantı tahmini performanslarını değerlendirmek için kesinlik ve *AUC* değerlendirme ölçütlerinin yanında daha adaletli bir değerlendirme ölçütü olan kesinlik-anma eğrileri de kullanılacaktır.



4. İKİ PARÇALI AKADEMİK BİLGİ AĞI İÇİN GELİŞTİRİLEN BAĞLANTI TAHMİNİ YÖNTEMİ

Bu bölümde büyük ölçekli gerçek dünya verilerinden oluşturulmuş iki parçalı akademik bilgi ağındaki yazarlarla konular arasındaki potansiyel bağlantıları tahmin etmek için güçlendirilmiş ağ modeline dayalı benzerlik tabanlı yeni bir tahmin edici önerilecektir. Önerilen yöntemin genel akış şeması Şekil 4.1’de verildiği gibidir. Bu çalışmanın iki önemli katkısı vardır. Bunlardan biri, konu tavsiyesi için büyük çaplı gerçek dünya verilerinden oluşturulmuş yazarlar ve çalıştıkları konu kümelerinden oluşan iki parçalı akademik bilgi ağının güçlendirilmiş tek parçalı ağlara dönüştürülmesi ve daha sonra hem iki parçalı ağ hem de güçlendirilmiş tek parçalı ağ bilgilerinin kullanmasıdır. Bu iki özelliğin birlikte kullanımına daha önce hiçbir çalışmada rastlanmamıştır. Diğeri ise, potansiyel bağlantıların kalitesini değerlendirmek amacıyla yeni bir yakınlık ölçütü algoritmasının önerilmesidir. Önerilen yakınlık ölçütü algoritmasının amacı, iki düğüm arasındaki bir bağlantının gücüyle ilgili öne sürülen Newman’ın [26] fikrini, iki düğüm arasındaki ortak komşuların sayısı ne kadar büyük olduğuyla orantılıdır, geliştirmektedir. Bu çalışmada güçlendirilmiş iki parçalı ağ kullanıldığı için, Newman’ın fikrini, düğümlerin ortak komşu sayıları, ortak komşularının ağ içerisinde ki önem dereceleri ve düğümlerin iki parçalı ağdaki önem derecelerinin göz önüne alındığı yeni bir bakış açısıyla değiştirdik. Önerilen yöntem gerçek dünya verileri üzerinde test edilecektir. Çalışmanın kilit noktalarından biri: gerçek dünya verilerinin toplanması ve metin analizi teknikleri kullanılarak iki parçalı yazar-konu ağında kullanılacak temel bilgilerin çıkartılmasıdır. Bu bölümde öncelikle veri toplama ve temel bilgi çıkarımından bahsedilecektir. Daha sonra önerilen temel bilgi çıkarma teknikleri kullanılarak işlediğimiz veri kümesiyle iki parçalı yazar-konu ağı oluşturulacaktır. Ağ oluşturulduktan sonra önerilen yöntemin detayları verilecektir. Önerilen yöntemin deneysel sonuçları, Intel Core i7 işlemciyle Windows 10 işletim sistemine ve 12 GB hafızaya sahip bir bilgisayarda gerçekleştirilmiştir. Veri tabanı olarak MySQL veri tabanı yönetim sistemi kullanılmıştır. Tez çalışmamızın hiçbir aşamasında hazır sosyal ağ analiz aracı kullanılmamıştır.



Şekil 4.1. Önerilen yöntemin genel mimarisi

4.1. Veri Toplama ve Temel Bilgi Çıkarımı

İki parçalı karmaşık ağlar için geliştirdiğimiz tahmin edici yöntemi test etmek için deneylerde gerçek dünya veri kümesi kullanılacaktır. Bu amaçla, bağlantı tahmini ile ilgili yapılan önceki çalışmalarda kullanılan veri kümelerinden farklı olarak IEEE Xplorer çevrimiçi kütüphaneden yazar-konu iki parçalı ağı oluşturmak için gerçek veri toplanmıştır.

IEEE(Elektrik ve Elektronik Mühendisleri Enstitüsü) Xplorer Dijital Kütüphanesi (<http://ieeexplore.ieee.org>): Ağırlıklı olarak bilgisayar bilimleri ve elektrik elektronik mühendisliği alanındaki bilimsel yayınların bilgilerinin tutulduğu güçlü bir kaynaktır. Bu kaynakta 4,1 milyondan fazla tam metin belge bulunmaktadır. İki milyondan fazla belge, sağlam, dinamik HTML biçimindedir. 174 üstünde IEEE dergi, yaklaşık 2000 e-kitap, yıllık olarak eklenen 1.400'den fazla IEEE konferansı bulunmaktadır. IEEE Xplorer'a her ay yaklaşık 20.000 yeni belge eklenmektedir.

IEEE Xplorer çevrimiçi kütüphaneden önceden seçilmiş bilgisayar bilimleri alanında düzenlenen 380 konferanstan 2000-2016 yılları arasında yayınlanmış 29.432 bildiri

toplanmıştır. Toplanan verilerin yıllara göre konferans ve bildiri sayısı dağılımı Tablo 4.1’de verilmiştir.

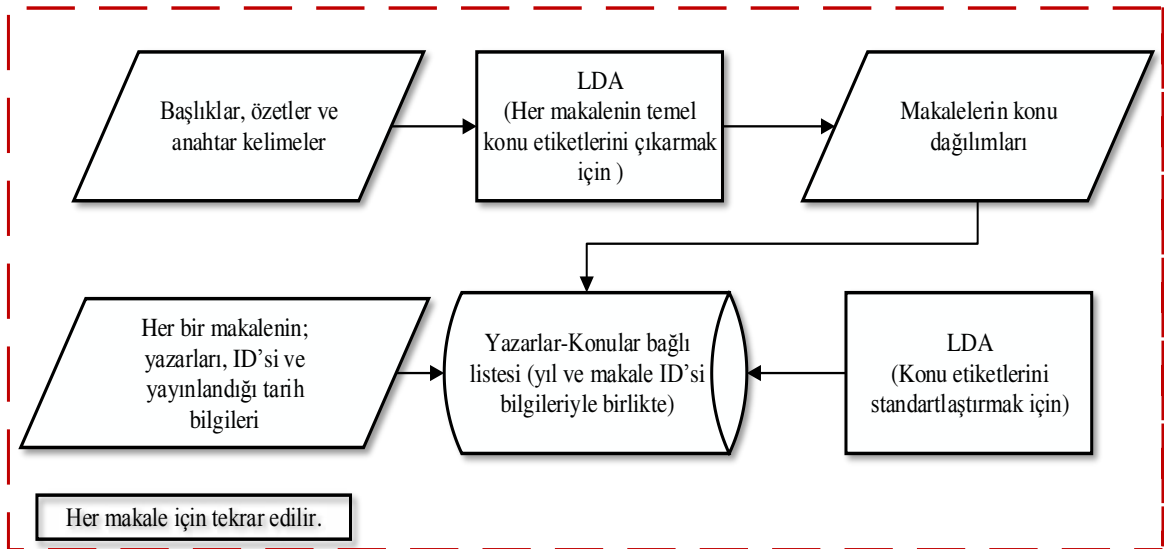
Tablo 4.1. Yıllara göre konferans ve bildiri sayısı

Yıl	Konferans Sayısı	Bildiri Sayısı
2000	12	371
2001	13	905
2002	10	505
2003	10	921
2004	14	1261
2005	20	1653
2006	24	1494
2007	20	1973
2008	13	1450
2009	19	1681
2010	22	2105
2011	26	2314
2012	29	2317
2013	29	2067
2014	31	2485
2015	39	2258
2016	52	3999
Toplam:	380	29432

Her bir bildiri için erişime açık “yazar isimleri, makale başlığı, anahtar kelimeler, özet ve referans listesi” gibi meta veriler toplanmış ve veri tabanına kaydedilmiştir. Toplanan bilgiler hiçbir ticari amaç olmaksızın sadece akademik çalışma içerisinde kullanılmıştır. Toplanan veri ham veridir. Verinin herhangi bir önışleme yapılmaksızın kullanılması eksik, gereksiz veya hatalı bilgiler içeren iki parçalı bir ağ oluşturulmasına neden olacaktır. Bu durum deneysel çalışmaların performans sonuçlarını olumsuz etkileyeceğinden dolayı bu çalışmada öncelikle ham veriden temel bilgilerin çıkarılması için Şekil 4.2’de gösterilen temel bilgi çıkarma mimarisi önerildi.

Bir makalenin içeriğini en doğru şekilde temsil edecek temel konu listesini bulmak amacıyla Şekil 4.2’de görüldüğü gibi özetler, anahtar kelimeler ve makale başlıkları olmak üzere üç bölüm kullanılır. Her bir bölüm kendi içerisinde makalenin çalışma konusunu temsil

etme yeterliliğinde anahtar sözcükler içerir. Üç bölümde de temel konu listesine aday olacak birbirine benzer sözcükler bulunur. Buradaki amaç, üç bölümdeki anahtar sözcükler arasındaki benzerliği bulan bir algoritma aracılığıyla kaliteli bir temel konu listesi çıkartmaktır. Bu amaçla, iki kelime arasındaki benzerliği derecelendiren *LDA* algoritması (Levenshtein Distance Algorithm) kullanılarak her bir makale için temel konu listesi çıkartılır ve ilgili makalenin her bir yazarı çıkartılan temel konu listesindeki konu etiketleri ile ilişkilendirilir. Daha sonra tüm konu etiketlerinin ve yazarların bulunduğu veri tabanında geçen konu etiketleri yeni bir listede birleştirilir. Bu liste *LDA* algoritmasına yeniden girdi olarak uygulanır. Son olarak konu listelerindeki her bir konu etiketi için standart bir etiket bulunur ve bunun paralelinde veri tabanındaki yazar-konu bağlantıları standartlaşmış konu etiketlerine göre güncellenir. Örneğin, veri madenciliği konu etiketi bir makalede “veri madenciliği” olarak geçerken başka makalede “veri madenciliği teknikleri” olarak geçmektedir. Başka bir örnek verirsek sosyal ağ analizi konu etiketi bir makalede “sosyal ağ analizi” olarak geçerken başka makalede “sosyal ağ analiz teknikleri” ya da “SAA teknikleri” olarak geçmektedir. Temel bilgi çıkarma tekniğimizle bu problemin önüne geçerek benzer konu etiketleri standart hale getirilir. Bu sayede gerçek dünya verilerinden oluşturulmuş ham veri tabanındaki bilgiler işlenir. İşlenmiş bu bilgilerle deneysel sonuçlarda kullanılacak iki parçalı akademik bilgi ağı modellenir.



Şekil 4.2. Temel bilgi çıkarma mimarisi

4.3. Yazar-Konu Akademik Bilgi Ağında Konu Tavsiyesi için Önerilen Bağlantı Tahmini Yöntemi

Bu bölümde, iki parçalı ağlarda potansiyel bağlantıları tahmin etmek için benzerlik tabanlı yeni bir bağlantı tahmini yöntemi öneriyoruz. Öncelikle büyük veriden Bölüm 4.1’de önerilen temel bilgi çıkarma yöntemiyle elde edilen işlenmiş değerli bilgiler kullanılarak Tanım 1’de verilen iki parçalı ağ oluşturulur.

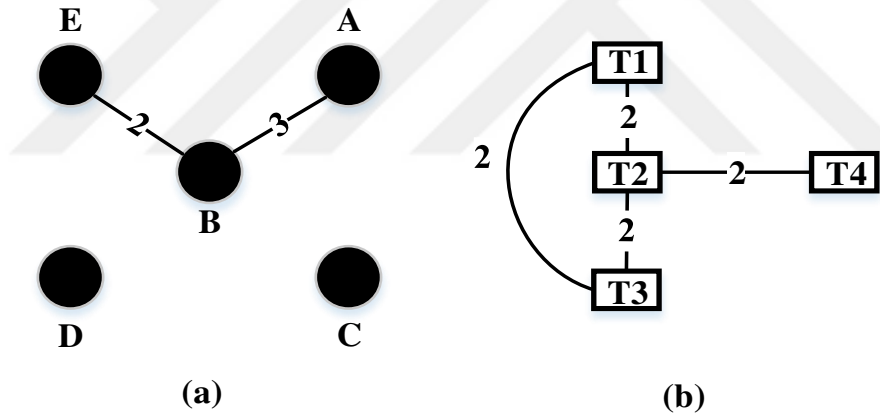
İki parçalı ağlarda bağlantı tahminiyle ilgili yapılan önceki çalışmaların neredeyse tamamı, düğüm çiftleri arasındaki bağlantıları tahmin etmek için iki parçalı ağlardan yansıtılmış tek parçalı ağların özelliklerini kullanmaya odaklanmıştır. İki parçalı ağlar ağırlıklı ve ağırlıksız olmaz üzere iki çeşit tek parçalı ağa dönüştürülebilir. Bu alanda yapılan çalışmalar incelendiğinde tek parçalı ağlardaki kenar ağırlıklarının özellikle önem taşıdığı görülür. Ağırlıklı ağlar analiz edilirken güçsüz bağlantı ve güçlü bağlantı kavramları ortaya çıkmıştır. Bu farkındalık, ağırlıklı ağlardaki güçsüz olarak nitelendirilecek bağlantıları tespit edip, mevcut ağdan daha güçlü bir ağ çıkarmayı amaçlayan yöntemlerin geliştirilmesine yol açmıştır. Bu tür güçlü ağ çıkarma yöntemleri, orijinal, ağırlıklı ağın sadece ağırlıklarını yeterince önemli gördüğü kenarları koruyan daha basit, ikili bir ağa indirgemeyi amaçlar [105]. Buradaki en büyük zorluk ağda önemli olarak nitelendirilebilecek kenar ağırlığının tespit edilmesidir. Bununla birlikte iki parçalı ağlar tek parçalı ağlara dönüştürüldüğünde bilgi kaybı olduğundan bağlantı tahmini yapılırken hem iki parçalı ağ hem de bu ağdan elde edilen güçlendirilmiş tek parçalı ağ bilgilerinin kullanılması özel bir dikkat gerektirir. Bu duruma dikkat eden yeni bir yöntem sunuyoruz.

Çalışmada gerçek dünya verilerinden elde edilmiş bir ağ kullanıldığından, ağ içerisinde ağın omurgası niteliğinde önemli yapılar olduğu kadar; önemsiz, ağın gelecek yapısıyla ilgili gereksiz bilgi taşıyan ilişkilerde bulunmaktadır. İki parçalı ağlarda bağlantı tahmini problemlerinde gereksiz ilişkiler gereksiz potansiyel bağlantıların oluşmasına sebep olur. Bu sebeple öncelikle ağın gereksiz ilişkilerden filtrelenmesi gerekir. Bu filtreleme sayesinde ağ karmaşıklığı azaltılarak daha kaliteli tahminler yapılması hedeflenir. Bu amaçla çalışmada, önceden yapılmış çalışmalardan farklı olarak ağırlıklı tek parçalı ağlar, omurga çıkarma yöntemi kullanılarak güçlendirilmiş tek parçalı ağlara dönüştürülür. Literatürde yapılan omurga özütleme çalışmaları incelendiğinde geliştirilmiş birkaç alternatif özütleme yöntemi olduğu görülür [113,114]. Bu çalışmada en çok kullanılan özütleme yöntemlerinden şartsız

eşik değeri belirleme yöntemi kullanılarak [12], iki parçalı yazar-konu akademik bilgi ağından yansıtılmış ağırlıklı tek parçalı yazar-yazar ağında maksimum gözlemlenen kenar ağırlığı güçlendirmede kullanılacak eşik değeri (α) olarak belirlenir.

Önerilen ağ modeli matematiksel olarak Tanım 5’te verildiği gibidir.

Tanım 5 (Güçlendirilmiş Tek Parçalı Ağ- (SBP^α)): Verilen $G = (U, V, E)$ iki parçalı ağdan kendini oluşturan her bir U ve V düğüm kümelerinden, U -yansıtılmış ağ $G_U = (U, E_U)$ ve V -yansıtılmış ağ $G_V = (U, E_V)$ olmak üzere iki tane tek parçalı ağ elde edilir. $(A, B) \in E_U, W: (A, B) \rightarrow |\Gamma(A) \cap \Gamma(B)|$. Eğer $W(A, B) > \alpha$ ise (A, B) kenarı SBP^α güçlendirilmiş ağda tutulurken, $W(A, B) < \alpha$ ağırlıkları eşik değerinin altında olan kenarlar güçlendirilmiş ağdan çıkartılır (α : önceden belirlenmiş eşik değeri). Bu şekilde güçlü düğüm çiftlerinin varlığını koruduğu ve güçsüz düğüm çiftlerinin yok olduğu tek parçalı ağın omurgası niteliğinde güçlendirilmiş bir ağ elde edilir. Bu durum için bir ağırlık fonksiyonu kullanılarak daha somut analizler yapılabilir. Bu durum Şekil 4.4’te açıklanmıştır.



Şekil 4.4. (a) Şekil 2.8(b)’de verilen yazar-yazar tek parçalı ağın $\alpha = 1$ için güçlendirilmiş yansıma ağı (SBP^1_{yazar}), (b) Şekil 2.8(c)’de verilen konu-konu tek parçalı ağın $\alpha = 1$ için güçlendirilmiş yansıma ağı (SBP^1_{konu}).

Önerilen yöntemde, gelecekte oluşması muhtemel potansiyel bağlantıları tespit etmek için SBP^α , de güçlü ilişkilerle bağlı düğüm çiftleri seçilir. Potansiyel bağlantılar için ağdaki tüm düğümlere odaklanmak yerine burada seçilen önemli düğüm çiftlerinin iki parçalı ağ bilgileri kullanılır. Bu sayede öngörülen bağlantı sayısı azalmakta, buna paralel olarak tahminlerin kalitesi de artmaktadır. İki parçalı ağı oluşturan düğümler alt ve üst düğümler olarak düşünülürse, G iki parçalı ağda var olan iki alt düğüm arasında hâlihazırda bir etkileşim varsa muhtemelen gelecekte de bu iki düğüm arasında etkileşim olacaktır. Öte yandan, G' ’de ortak

bir komşusu olmayan iki alt düğümün gelecekte etkileşim içerisine girmesi düşük bir ihtimaldir. Hâlihazırda ağırlıklı tek parçalı ağdan elde edilmiş güçlendirilmiş tek parçalı ağda (SBP^α) birbirlerine güçlü ilişkilerle bağlı düğüm çiftleri arasında iki parçalı ağda "iç düğümler" olduğunu varsayarsak yapılan tahminlerin doğruluk oranları daha yüksek olacaktır.

Tanım 6 (SBP^α ile Potansiyel Bağlantılar (PL)): $G = (U, V, E)$ iki parçalı ağ, $G_U^\alpha = (U, E_U^\alpha)$, α eşik değeriyle güçlendirilmiş U - tek parçalı ağ ve aynı şekilde $G_V^\alpha = (V, E_V^\alpha)$, V - güçlendirilmiş tek parçalı ağdır. G ağından elde edilen U -yansıtılmış $G_U^\alpha = (U, E_U^\alpha)$ ağının elemanlarından $A \in U$ için $k_i \in \Gamma_A^\alpha$, $i = 1, \dots, l$, olduğunu varsayarsak A düğümü için gelecekte oluşma ihtimali olan potansiyel bağlantılar (PL) bulunurken aşağıda verilen Denklem (4.1) temel alınır.

$$PL_A = \{K \setminus (\Gamma_A)_{eğitim}\}, K = \{\Gamma(k_1) \cup \Gamma(k_2), \dots \cup \Gamma(k_l)\}, \quad (4.1)$$

Denklem (4.1)'de, $(\Gamma_A)_{eğitim}$ A düğümünün $G_{eğitim}$ ağındaki komşularını ve Γ_A^α ise A düğümünün G_U^α yansıtılmış ağdaki komşularını temsil etmektedir ($k_i \in \Gamma_A^\alpha$, $i = 1, \dots, l$).

K kümesindeki her bir $\Gamma(k_l)$, k_l düğümünün G iki parçalı ağdaki komşularını temsil eder.

$U = \{A, B, \dots, m\}$ için gelecekte oluşması öngörülen bağlantılar $PL = \{PL_A \cup PL_B, \dots, \cup PL_m\}$ 'dir. Benzer şekilde, potansiyel bağlantılar (PL), V –yansıtılmış ağ içinde tanımlanabilir. Hesaplama süresini azaltmak için iki parçalı ağı sadece daha az düğümlü parçaya göre yansıtmak yeterlidir. Örneğin, $|U| < |V|$ için U –yansıtılmış ağ hesaplama karmaşıklığını azaltmak için kullanılabilir.

Tanım 6'dan yararlanılarak elde edilen (PL) kümesinde ki her bir elemanın potansiyel bağlantı statüsünde olması için aşağıdaki şartı sağlaması gerekir.

(A, p_i) , $G = (U, V, E)$ iki parçalı ağdaki bir düğümün U –yansıtılmış potansiyel bağlantı olması için gerekli şart:

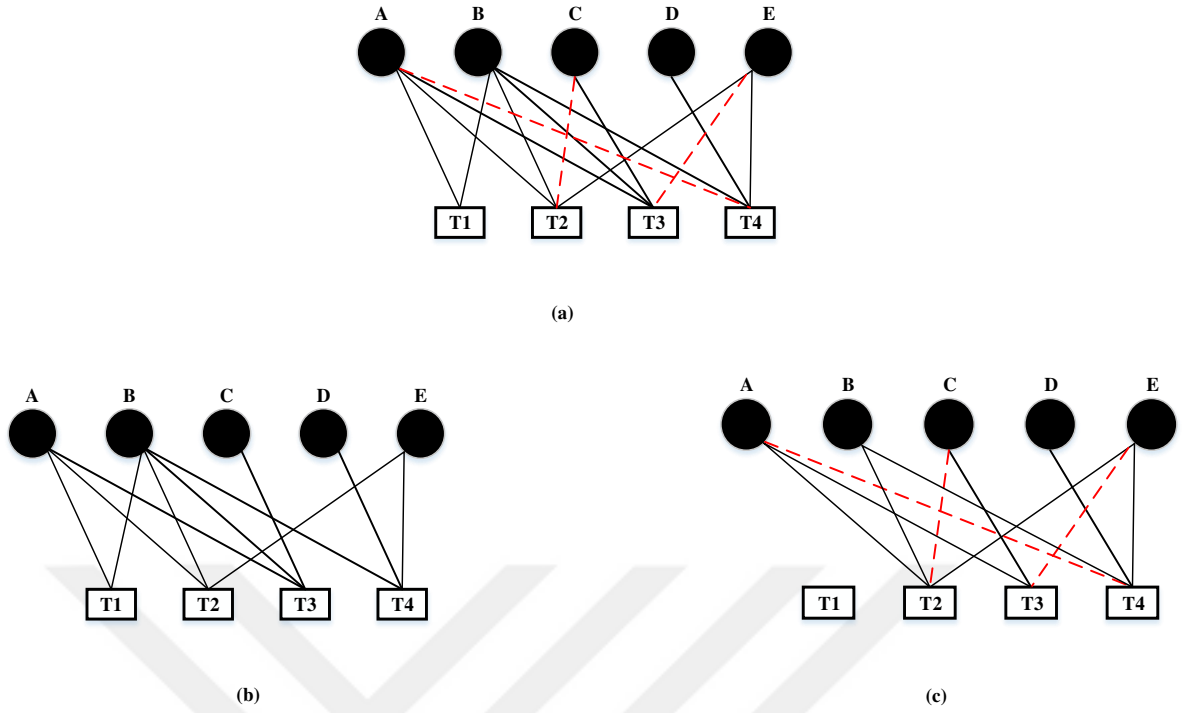
$$\Gamma_U(A) \cap \Gamma(p_i) \neq \emptyset \text{ ve } p_i \notin \Gamma(A), \emptyset \text{ boş kümedir.} \quad (4.2)$$

$A \in U$ ve $p_i \in V$, $(A, p_i) \notin E$ iki düğümdür, $(A, p_i) \in U \times V$ yeni düğüm çifti aslında iç bağlantı özelliğine sahiptir. İç bağlantı olma özelliğinin detayları Bölüm 3'te Tanım 4'te detaylı olarak açıklanmıştır.

Tanım 7 (Potansiyel Bağlantıların Kapsadığı Örüntü): (A, p_i) , $G_{eğitim}$ iki parçalı eğitim ağı, G_U^α ise $G_{eğitim}$ ağının U parçasından α eşik değeriyle yansıtılmış, güçlendirilmiş tek parçalı ağıdır. Her bir $C \in (\Gamma_U(A))_{eğitim} \cap \Gamma(p_i)$ için $\{A, C\}$, (A, p_i) potansiyel bağlantı (PL) tarafından kapsanan örüntüdür.

Bir PL 'nin kapsadığı örüntü sayısı, kendisini oluşturan kenar düğümlerin iki parçalı ağda birbirleriyle olan yakınlıklarıyla doğrudan alakalıdır. PL 'nin kapsadığı örüntü sayısı ne kadar fazlaysa, gelecekte oluşma ihtimali de o kadar fazladır. Bu yüzden PL tarafından kapsanan örüntü sayısı potansiyel bağlantının tahmin olasılığını ölçmek için kullanılabilir.

Örneğin, Şekil 4.5(a)'da verilen iki parçalı ağda üstteki dairesel düğümler yazarları alttaki kare düğümler ise yazarların çalıştığı konu etiketleridir. (E, T_1) ve (E, T_3) bağlantıları Tanım 6 şartını sağlayan PL kümesine ait potansiyel bağlantılardır. Bu iki PL 'nin kapsadığı örüntü sayısı farklıdır. (E, T_1) bağlantısının kapsadığı örüntü $\{E, A\}, \{E, B\}$ iken, (E, T_3) bağlantısı tarafından kapsanan örüntü $\{E, A\}, \{E, B\}, \{E, C\}$ 'dir. Önerilen yöntemle göre, gelecekte oluşması öngörülen (E, T_3) potansiyel bağlantının oluşma ihtimali (E, T_1) potansiyel bağlantısının oluşma ihtimalinden daha fazladır. Bu sebeple potansiyel bağlantıların skorları hesaplanırken kapsadığı örüntü sayısını temel alınır.

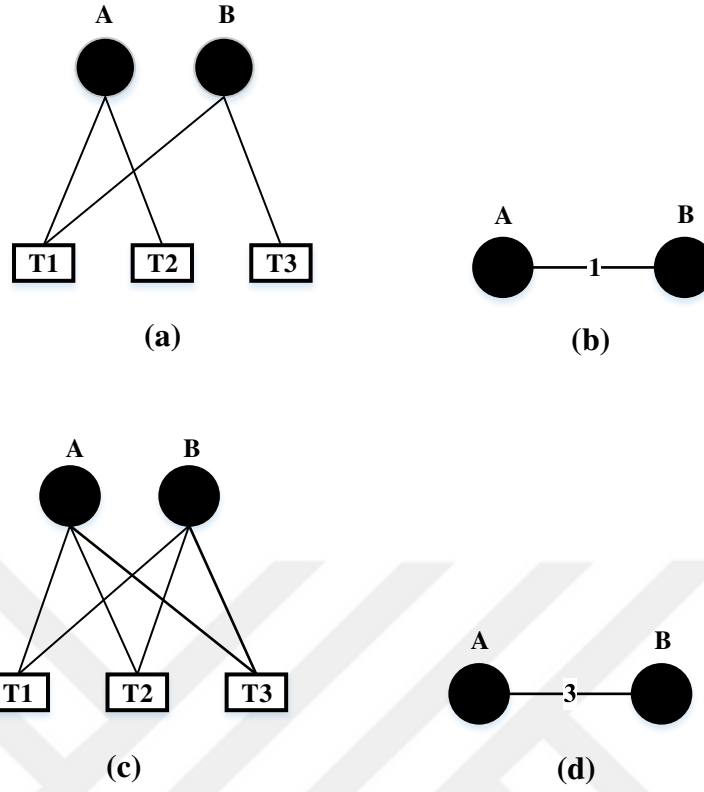


Şekil 4.5. (a) İki parçalı yazar-konu ağı, (b) İki parçalı yazar-konu eğitim ağı, (c) İki parçalı yazar-konu test ağı.

Tanım 8 (Örüntü Ağırlığı): PL 'nin kapsadığı örüntü sayısı ne kadar önemliyse kapsanan örüntüde ki her bir bağlantının ağ içerisinde ki ağırlığı da bir o kadar önemlidir [27]. $\{A, B\}$ örüntü ağırlığını hesaplarken üç temel faktör dikkate alınır:

(1) A ve B düğümlerinin iki parçalı ağdaki ortak komşu sayıları

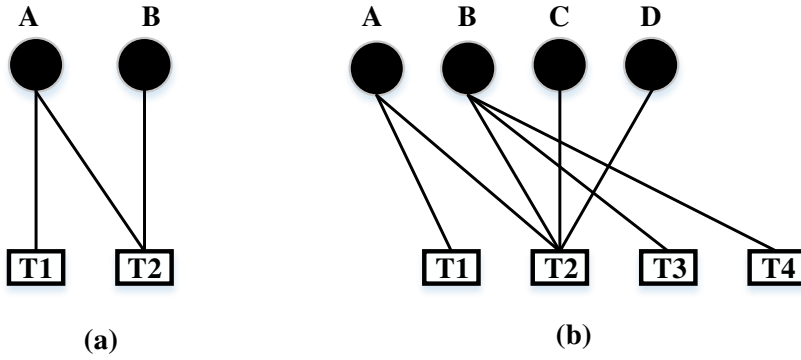
Bir potansiyel bağlantının (PL) kapsadığı örüntü yansıtılmış ağdaki kenarlardır. G iki parçalı ağda ortak komşu düğümlere sahip A ve B kenarları, G_U yansıtılmış ağda (A, B) kenarıyla temsil edilir. İki parçalı ağlar tek parçalı ağlara dönüştürülürken bazı topolojik bilgiler kaybolur. Şekil 4.6(a) ve Şekil 4.6(c) de verilen iki parçalı ağları birbirlerinden farklı olmasına rağmen yansıtılmış ağları Şekil 4.6(b) ve Şekil 4.6(d)'de görüldüğü gibi aynıdır. Böyle durumlarda iki parçalı ağın yapısındaki topolojik bilgilerde ciddi bir kayıp meydana gelecektir. Bu problemle başa çıkmak, ağ bilgilerini bir nebze de olsa korumak amacıyla ağırlıklı tek parçalı ağlar oluşturulur. Şekil 4.6(a)'da A ve B düğümlerinin ortak komşu sayıları Şekil 4.6(b)'de ki komşu sayısından küçük olduğu için daha küçük derecelendirilir.



Şekil 4.6. İki farklı iki parçalı ağ ve bunların üstteki düğüm kümesinden yansıtılmış ağırlıklı tek parçalı ağ. (b), (a)'da verilen iki parçalı ağdan yansıtılmış tek parçalı ağırlıklı ağ. (d), (c)'de verilen iki parçalı ağdan yansıtılmış tek parçalı ağ.

(2) A ve B düğümlerinin iki parçalı ağdaki ortak komşularının dereceleri

$\{A, B\}$ potansiyel bağlantısının iki parçalı ağdaki ortak komşu sayılarının yanı sıra ortak komşularının iki parçalı ağdaki dereceleri de önemlidir. İki düğümün ortak komşularının iki parçalı ağdaki dereceleri arttıkça bu iki düğümün birbirlerine olan benzerlikleri azalacaktır. Örneğin, Şekil 4.7(a)'da A ve B yazarlarının ortak çalışma konusu olan T2 düğümünün derecesi ikidir, Şekil 4.7(b)'de ise yine tek bir ortak çalışma konusu olan T2 konusunun derecesi dörttür. Şekil 4.7(a)'da yalnızca A ve B düğümünün çalıştığı T2 konusu yüksek akademik benzerlik ifade ederken Şekil 4.7(b)'de dört kişi tarafından çalışılan T2 konusu A ve B düğümleri için daha düşük akademik benzerlik ifade eder. Sonuç olarak iki düğümün ortak komşularının dereceleri ne kadar küçükse $\{A, B\}$ benzerliği o kadar fazladır ve örüntü ağırlığı o kadar yüksektir.



Şekil 4.7. İki farklı iki parçalı ağ.

(3) A ve B düğümlerinin iki parçalı ağdaki dereceleri

A ve B düğümlerinin iki parçalı ağdaki derecelerinin $\{A, B\}$ örüntü ağırlığına negatif bir etkisi vardır. Örneğin, Şekil 4.7(a)'da A ve B düğümlerinin dereceleri sırasıyla 2 ve 1'dir. Şekil 4.6(b)'de ise A ve B düğümlerinin dereceleri sırasıyla 2 ve 3'tür. Şekil 4.7(a)'da A yazarı T1 ve T2 konularına çalışmışken B yazarı sadece T2 konusuna çalışmıştır. Fakat Şekil 4.7(b)'de farklı birçok alanda çalışma yapmış olup tek ortak çalışma alanları T2'dir. Bu yüzden Şekil 4.7(b)'deki akademisyenlerin çalışma alanları Şekil 4.7(a)'daki akademisyenlerin çalışma alanlarına göre daha az benzerlik taşımaktadır, haliyle $\{A, B\}$ ağırlığı Şekil 4.7(a)'da daha yüksek olmalıdır.

Örüntü Ağırlığı: $G_U = (U, E_U)$, $G = (U, V, E)$ iki parçalı ağdan elde edilmiş yansıtılmış tek parçalı ağırlıklı ağdır. $(A, B) \in E_U$, G_U ağında bir kenardır. Bu kenarın ağırlığı aşağıda seçilen algoritmalar ile hesaplanır:

Önerilen yakınlık ölçütü algoritmasıyla potansiyel bağlantı skoru aşağıdaki gibi hesaplanır:

$$w(A, B) = \frac{2}{k(A) + k(B)} \sum_{z \in \Gamma(A) \cap \Gamma(B)} \frac{1}{k(z)} \quad (4.3)$$

Burada, k_A , k_B ve k_z sırasıyla A, B ve z ($A, B \in U, z \in V$) düğümlerinin $G = (U, V, E)$ iki parçalı ağdaki düğüm dereceleridir, diğer bir deyişle komşu sayılarıdır. $\Gamma(A), \Gamma(B)$ ise sırasıyla A ve B düğümlerinin iki parçalı ağdaki komşularıdır. Denklemden de anlaşıldığı gibi A ve B düğümlerinin ortak komşularının dereceleri ne kadar küçükse örüntü ağırlığı o kadar fazladır. Bir PL'nin kapsadığı her bir örüntü elemanının oluşma olasılığı kapsadığı

örüntülerin toplam ağırlıklarına eşdeğerdir. $G_{eğitim} = (U, V, E_{eğitim})$ ağında ilerde oluşması öngörülen potansiyel bağlantılardan (A, p_i) potansiyel bağlantısının final skoru aşağıdaki gibidir:

$$S(A, p_i) = \sum_{\{A, B\} \in \Gamma(A, p_i)} w(A, B) \quad (4.4)$$

Denklemden de görüldüğü gibi eğer (A, p_i) potansiyel bağlantı ne kadar yüksek ağırlıkta örüntü kapsıyorsa, o kadar yüksek bağlantı tahmini olasılığına sahiptir. Bu yüzden bir potansiyel bağlantının kapsadığı her bir örüntü ağırlığının toplamı, o potansiyel bağlantının bağlantı tahmini final puanıdır.

4.4. Önerilen Yöntemin Sözde Kodu

Giriş: $G(U, V, E)$ iki parçalı ağ, $G_{eğitim}(U, V, E_{eğitim})$ iki parçalı eğitim ağı, α:eşik değeri; Çıkış: (PL) potansiyel bağlantı kümesi ve bu kümedeki her bir elemanın bağlantı tahmini final skoru.	
Part 1: Tanım 3'e göre G_u yansıtılmış ağırlıklı ağ oluşturulması <pre> 1. for iki parçalı ağdaki $\forall A \in U$ do 2. for $\forall x \in \Gamma(A)$ do 3. for $\forall B \in \Gamma(x)$ do 4. $W: (A, B) \rightarrow \Gamma(A) \cap \Gamma(B)$ 5. $E_U = E_U \cup \{(A, B)\}$ 6. end for 7. end for 8. end for </pre>	Part 2: Tanım 5'e göre G_u yansıtılmış ağın G_U^α güçlendirilmiş tek parçalı ağa dönüştürülmesi <pre> 9. for $\forall A \in G_U$ do 10. for $\forall B \in \Gamma_U(A)$ do 11. if $W(A, B) > \alpha$ then 12. $E_U^\alpha = E_U^\alpha \cup \{(A, B)\}$ 13. end if 14. end for 15. end for </pre>
Part 3: Tanım 6'ya göre PL (Potansiyel Bağlantılar) kümesinin çıkartılması <pre> 16. for G_U^α güçlendirilmiş tek parçalı ağdaki $\forall A \in U$ do 17. PL_A boş dizi 18. for $\forall K \in \Gamma_U^\alpha(A)$ do 19. for $\forall p \in \Gamma(K)$ do 20. if $(A, p) \notin E_{eğitim}$ then 21. $PL_A = PL_A \cup (A, p)$ 22. end if 23. end for 24. end for 25. $PL = PL \cup PL_A$ 26. end for </pre>	Part 4: PL kümesinin her bir elemanının bağlantı tahmini final skorunun hesaplanması <pre> 27. for $\forall (A, p) \in PL$ do 28. for $\forall C \left(C \in (\Gamma_u(A))_{train} \cap \Gamma(p) \right)$ do 29. (A, p) potansiyel bağlantı için Denklem (4.3)'e göre (A, C) kapsanan örüntü ağırlığını hesapla 30. Denklem (4.4)'e göre 31. $S(A, p) = S(A, p) + w(A, C)$ 32. end for 33. end for </pre>

Allali ve ark. [54], iki parçalı ağlarda iç bağlantılar kavramını tanımlamış ve daha sonra bu tanıma dayalı bir bağlantı tahmini yöntemi önermiştir. Gao ve ark. [11] ise [106]'da tanımlanan iç bağlantılar yöntemine dayalı aday düğüm çiftleri kavramını tanımladı ve buna dayalı bir bağlantı tahmini yöntemi sundu. Bu çalışmada ise her iki yöntemin özelliklerini geliştirerek gerçek dünya verilerinden oluşturulmuş büyük çaplı iki parçalı ağlarda bağlantı tahmini problemine uygun yeni bir yöntem önerdik. Önerdiğimiz yöntemin [106] ve [11]'de sunulan yöntemlerle benzer olmasının yanında önemli bazı farklılıkları da vardır. [106]'da yansıtılmış ağın iki parçalı ağın herhangi bir parçası üzerinde gerçekleştirebileceği ifade edilmiş, ancak farklı yansıtılmış ağların farklı sonuçlara neden olup olmayacağı ile ilgili herhangi bir açıklama yapılmamıştır. [11]'de ise bu duruma açıklık getirmek amacıyla bir iç bağlantının, ağın her iki yansıması içinde iç bağlantı olacağı ispatlanmış bu sebeple hesaplama süresini azaltmak için iki parçalı ağın sadece daha az düğümlü parçaya göre yansıtmanın yeterli olduğu vurgulanmıştır. Bu tez çalışmasında ise ağın az düğümlü parçasından elde edilmiş güçsüz ilişkilere sahip düğümleri içeren yansıtılmış tek parçalı ağ, ağın omurgası niteliğini taşıyan güçlendirilmiş tek parçalı ağa yansıtılır. Güçlendirilmiş bu ağ önceden belirlenmiş eşik değerine göre oluşturulur. Eşik değeri, başlangıçtaki tek parçalı yansıtılmış ağda en çok gözlemlenen kenar ağırlığının belirli bir yüzdesine göre belirlenir. Bu sayede hem potansiyel bağlantı kümesi azalacak ve hesaplama süresi kısılacak hem de en güçlü ilişkilerle ağdaki varlığını koruyan düğüm çiftleri arasındaki güçlü bağlantılar bulunacaktır. [106]'da potansiyel bağlantılar bulunurken mevcut ağda olmayan tüm bağlantılar ele alınır. [11]'de ise tek parçalı yansıtılmış ağdaki düğüm çiftleri arasındaki ilişkiler kullanılarak iki parçalı ağda olmayan fakat ileride oluşması öngörülen potansiyel bağlantılar tespit edilir. Önerdiğimiz yöntemde ise ağın omurgası niteliğini taşıyan, güçlendirilmiş ağırlıklı tek parçalı ağda varlığını koruyan düğüm çiftleri arasındaki ilişkiler kullanılarak eğitim ağında olmayan fakat gelecekte oluşması öngörülen potansiyel bağlantılar tespit edilir. Bu şekilde [11]'de ki güçsüz ilişkiye sahip düğüm çiftlerine önerilecek güçsüz potansiyel bağlantıları filtrelemiş oluruz. [11] ve [106]'da küçük veri ağları kullanıldığından potansiyel bağlantılar bulunurken düğüm çiftlerinin yalnızca eğitim ağındaki ilişkilerine bakılırdı. Önerdiğimiz yöntemle akademik bilgi ağında yazarlara konu önerme hedeflenir. Akademik dünyada konu başlıkları belirli yıllardan sonra popülerleşir. Bu sebeple potansiyel bağlantılar bulunurken bilgi kaybını önlemek amacıyla eğitim ağıyla güçlendirilmiş tek parçalı ağdan güçlü düğüm çiftleri bulunduktan sonra, düğüm çiftleri arasındaki eksik bağlantılar bulunurken konu önerilecek yazarın çalışma konuları eğitim ağından alınırken

komşularının çalıştığı konular ise test verisinin de eklendiği iki parçalı ağdan alınır. Böylelikle güncel çalışma konuları kaybedilmemiş olur.

4.5. Uygulama Sonuçları

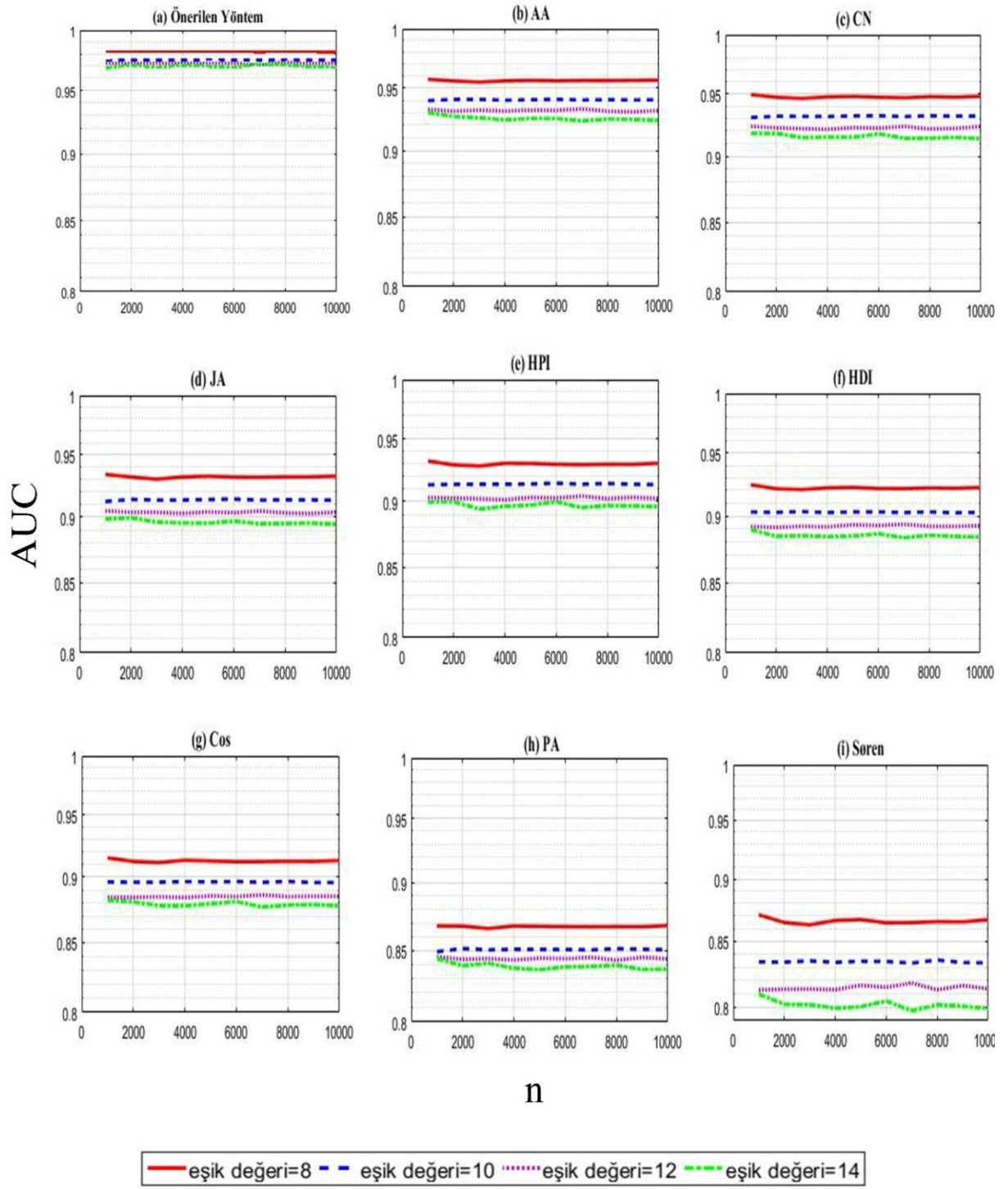
Deneylerimizin ilk aşamasında, Bölüm 4'te önerdiğimiz bağlantı tahmini yöntemi $\alpha = 8, \alpha = 10, \alpha = 12$ ve $\alpha = 14$ eşik değerleri kullanılarak test edildi. Önerilen yöntemde uygun eşik değerinin bulunması en önemli problemidir. Bu tez çalışmasında, uygun eşik değerini bulmak amacıyla öncelikle yazar-konu iki parçalı ağdan elde edilen yansıtılmış ağırlıklı tek parçalı ağdaki kenar ağırlıklarının frekansları hesaplanır. Daha sonra, frekans değeri en yüksek kenar ağırlığı en uygun eşik değeri olarak seçilir. Yapılan hesaplama sonucunda ağda frekans değeri en yüksek diğer bir deyişle en çok gözlemlenen kenar ağırlığı “10” olarak tespit edildi. Bu değer ardışıklarına ayarlanan dört farklı eşik değeri seçilerek dört ayrı bağlantı tahmini kümesi elde edildi ve sonuçlar dört kez test edildi. Seçilen eşik değerleri kullanılarak dört ayrı potansiyel bağlantı tahmin kümesi elde edildikten sonra, kümenin her bir elemanı için önerdiğimiz bağlantı tahmini algoritmasıyla tahmin skoru hesaplandı. Önerilen yeni benzerlik ölçütü algoritması, literatürde yaygın olarak kullanılan *AA, CN, JA, HPI, HDI, Cos, PA, Sørensen* benzerlik ölçütleriyle karşılaştırıldı. Karşılaştırma için kullanılan algoritmaların detaylı açıklaması Bölüm 3.2’de verilmiştir. Algoritmaların tahmin performansları *AUC, kesinlik* ve *kesinlik-anma* değerlendirme ölçütleri kullanılarak karşılaştırıldı.

Şekil 4.8 her bir algoritmanın dört ayrı eşik değerine göre uygulanan yöntemimizden elde edilen potansiyel bağlantı kümesinde ki *AUC* değerlendirme ölçütüyle hesaplanan tahmin performanslarını gösterir. Ayrıca her bir algoritmanın dört eşik değerine göre performansları da karşılaştırılır. *AUC* kalitesi test edilirken grafiklerde yatay eksen de gösterilen her bir n değerinin *AUC* skoru için rastgele n seçim 30 kez tekrarlanmıştır. *AUC* bağlantı tahmini kalitesini gösteren skor; her bir n değeri için hesaplanan 30 performans değerinin ortalamasıdır.

Şekil 4.8 incelendiğinde önerilen yöntemin dört eşik değeri için eksen grafiğinde daha dar bir alan kapladığı ve her bir eşik değeri için *AUC* kalitesinin 1’e daha yakın olduğu, diğer algoritmaların ise eksen grafiğinde kapladığı alanın genişlediği ve 1’den uzaklaştığı görülür. Tüm algoritmalar içinde dört versiyon üzerinde, tüm n değerleri için en iyi tahmin kalitesi önerdiğimiz yönteme aittir. Aynı zamanda Şekil 4.8’de görüldüğü gibi her bir algoritma için

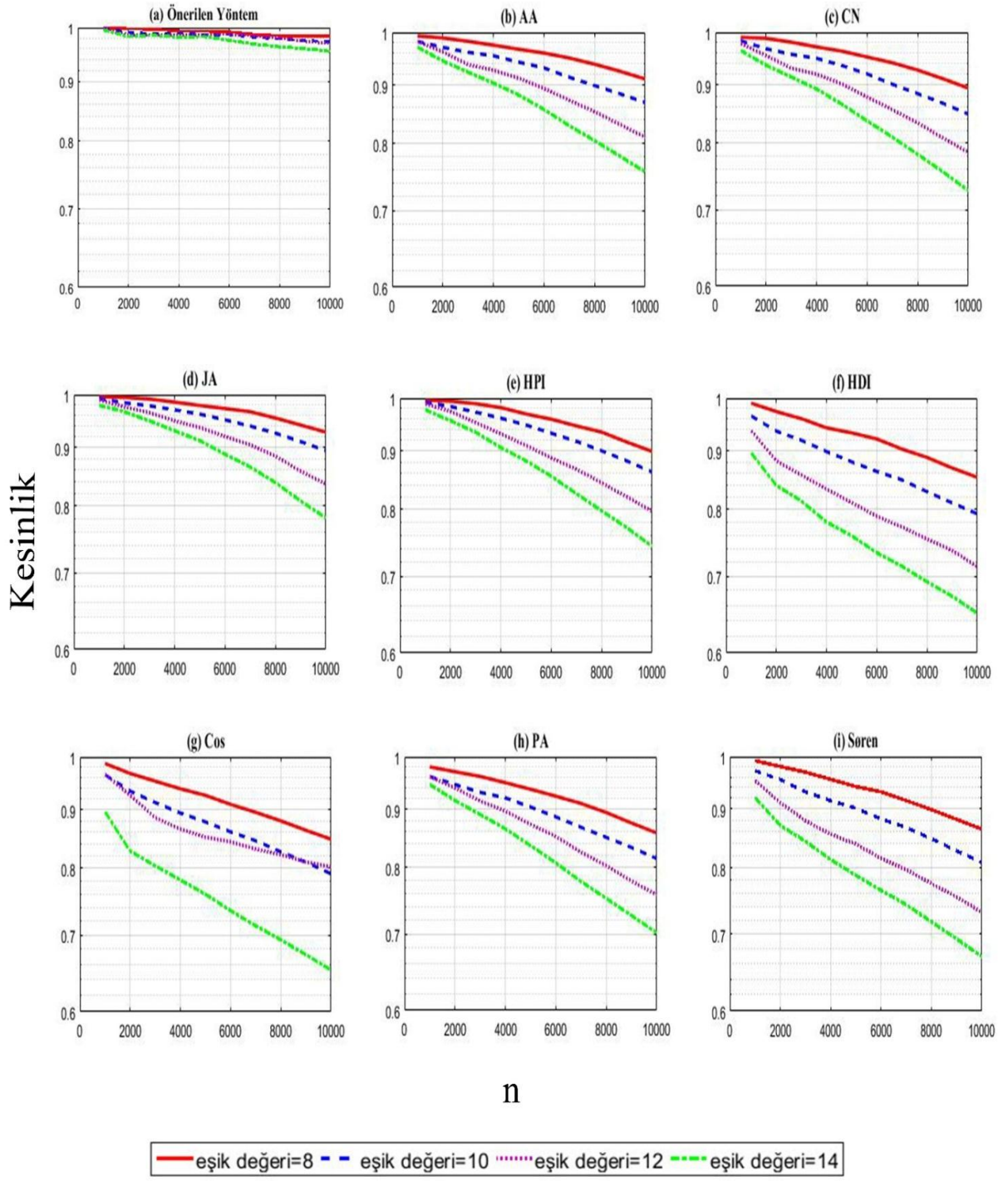
en kötü tahmin kalitesi $\alpha = 14$ eşik değeriyle elde edilen dördüncü versiyon bağlantı tahmini kümesinde (*PL*) dir.

Her bir algoritmanın performansı detaylı olarak incelendiğinde anlaşılırki, önerilen yöntem dört ayrı versiyon potansiyel bağlantı kümeleri içindeki performansları karşılaştırıldığında minimum değişimle diğer algoritmalarla karşılaştırıldığında en iyi sonuca sahiptir. Bu durumun en önemli nedeni, diğer algoritmalarla örüntü ağırlığı hesaplanırken düğüm derecesi ortak komşu vs. gibi özelliklerden sadece biri kullanılır. Önerilen yöntemde ise hem düğümlerin ortak komşu sayıları, hem ortak komşularının ağ içerisindeki önem dereceleri hem de düğümlerinin iki parçalı ağdaki önem dereceleri göz önüne alındığından ağın topolojik yapısındaki bilgi kaybı minimuma indirilmiştir. Şekil 4.8’den de anlaşıldığı gibi eşik değeri büyüse bile önerilen yöntem diğer algoritmalarla kıyasla bu bilgi kaybından en az etkilenmiştir. Örneğin, *Adamic-Adar* (denklem (3.3)) algoritması, önerilen yöntemdeki örüntü ağırlığı hesaplamasında kullanılan üç kriterden sadece biri olan “*ortak komşuların iki parçalı ağdaki dereceleri*” kriterini dikkate alır. Şekil 4.8’de görüldüğü gibi, en yüksek performansa ve minimum değişiklik miktarına sahip ikinci algoritma, *Adamic-Adar* algoritmasıdır. Her bir algoritma için en iyi tahmin kalitesi önerilen yöntemin $\alpha = 8$ eşik değeriyle elde edilen birinci versiyon bağlantı tahmini kümesinde (*PL*) elde edilmiştir.



Şekil 4.8. Algoritmaların farklı eşik değerlerindeki AUC bağlantı tahmini kalitelerinin kıyaslanması

Önerilen yöntemin performansını daha iyi anlamak amacıyla bağlantı tahmini kalitesini farklı bir perspektiften ölçen “kesinlik” değerlendirme ölçütü kullanıldı. Şekil 4.9, her bir algoritmanın dört ayrı eşik değerine göre uygulanan yöntemimizden elde edilen potansiyel bağlantı kümesinde ki kesinlik ölçütüyle hesaplanan tahmin performanslarını gösterir. Ayrıca her bir algoritmanın dört ayrı eşik değerlerine göre performansları da karşılaştırılmıştır. “AUC” bağlantı tahmini değerlendirme ölçütüyle benzer şekilde en iyi tahmin performansları tüm eşik ve n değerleri için önerdiğimiz yönetime aittir. Hem AUC hem de kesinlik grafiklerinde deneysel sonuçların test edildiği diğer algoritmalarla kıyaslandığında tüm eşik eğrilerinde her bir n değeri için en iyi tahmin kalitesi önerdiğimiz yönetime aittir. Her iki şekil incelendiğinde görülür ki algoritmaların bağlantı kaliteleri her iki değerlendirme ölçütü için $n = 1000$ ’de birbirine yakın değerlerken n değeri arttıkça iki grafik arasındaki farkta artmaktadır (Bkz. Şekil 4.8(a) ve Şekil 4.9(a)). AUC grafiklerinde her bir algoritmanın eşik değerlerindeki bağlantı tahmini kalitesinin n değerindeki artışlarda ihmal edilebilecek oranda değiştiği kesinlik grafiklerinde ise her bir algoritmanın eşik değerlerindeki bağlantı tahmini kalitesinin ihmal edilemeyecek oranda değiştiği gözlemlenir. Şekil 4.8’de görüldüğü gibi her bir eşik değeri için maksimum ve minimum kalite değerleri arasında büyük fark yokken kesinlik benzerlik ölçütünde bu fark artmaktadır. Örneğin, Sørensen algoritmasının tahmin kalitesi sonuçlarının AUC ölçütü ile hesaplanan maksimum ve minimum değerleri arasındaki fark 0,09 iken ($\alpha = 8, n = 1000$, maksimum= 0,8709, minimum= 0,79895), aynı noktalarda kesinlik tahmin grafiğinde bu fark 0,32’ye çıkmaktadır (maksimum=0,992, minimum= 0,6694) (Bkz. Şekil 4.8(i) ve Şekil 4.9(i)). Bu durumun temel sebebi, kesinlik ile tahmin kümesinden sadece *top – n* bağlantı tahmini sonucu kalitesi göz önüne alınırken, AUC ile algoritmanın tüm tahmin kümesi sonuçlarının kalitesi göz önüne alınmasıdır.

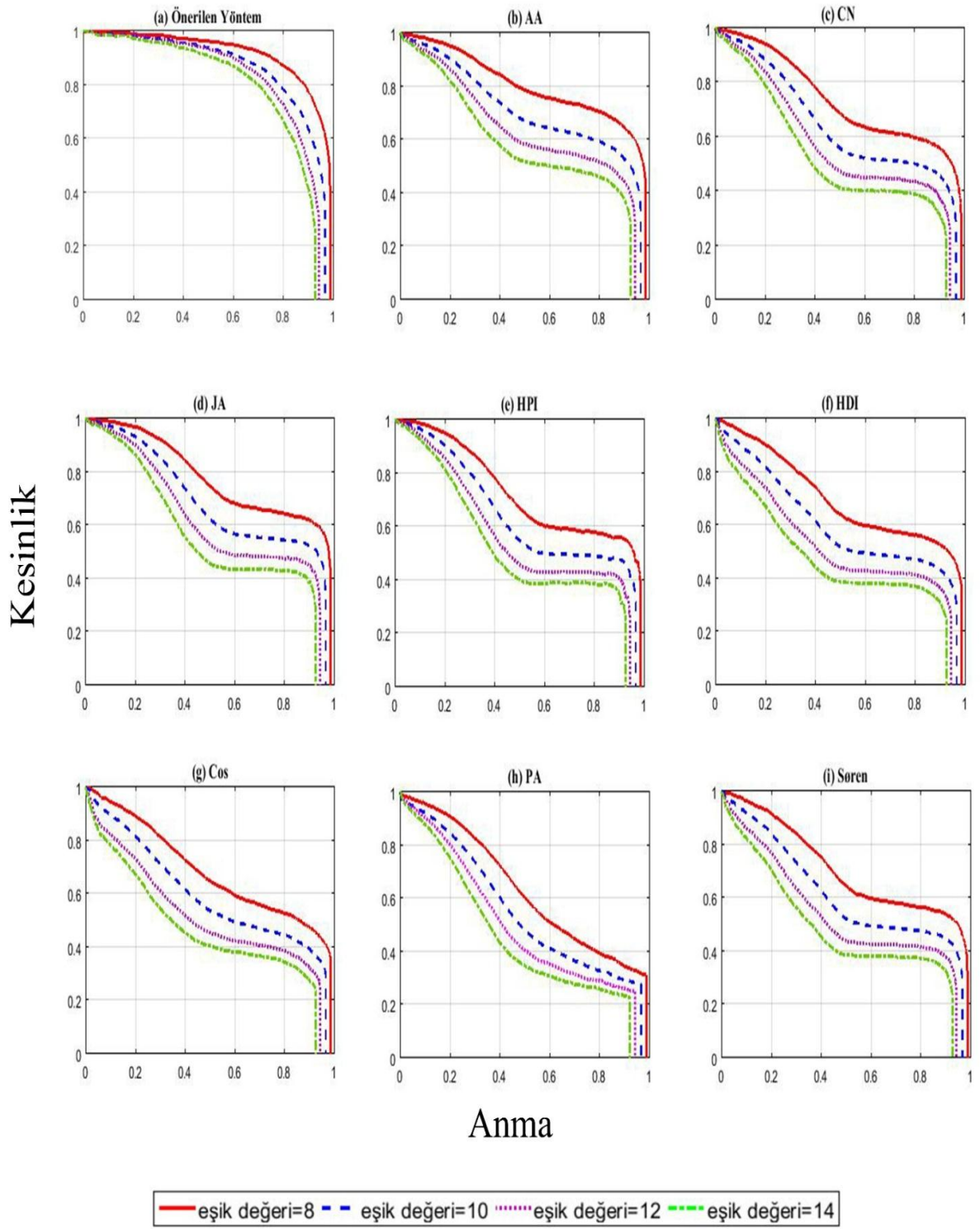


Şekil 4.9. Algoritmaların farklı eşik değerlerindeki kesinlik bağlantı tahmini kalitelerinin kıyaslanması

Tablo 4.2. Eşik değerlerine göre önerilen yöntemle elde edilen potansiyel bağlantı kümelerinin eleman sayıları

Eşik Değerleri	$\alpha = 8$	$\alpha = 10$	$\alpha = 12$	$\alpha = 14$
PL (Potansiyel Bağlantı Kümesi)	9.787.561	7.187.242	4.614.604	3.127.562

Şekil 4.8 ve Şekil 4.9’da görüldüğü gibi her bir algoritma için en iyi tahmin kalitesi, $\alpha = 8$ eşik değeriyle elde edilen birinci versiyon bağlantı tahmini kümesinden (PL) elde edilmiştir. Ayrıca, Şekil 4.8 ve Şekil 4.9’da görüldüğü gibi eşik değeri büyüdükçe algoritmaların tahmin kalitesi de düşmektedir. Fakat eşik değeri küçüldükçe de Tablo 4.2 de ifade edildiği gibi potansiyel bağlantı kümesi büyümekte buna paralel olarak algoritmanın çalışma zamanı ve tahmin kalitesi artmaktadır. Fakat yukarıdaki şekillerde de görüldüğü gibi bu durum aslında negatif etki yaratıp bilgi kaybına sebep olacaktır (bkz. Şekil 4.8, Şekil 4.9). Tüm algoritmalarda AUC tahmin kalitesi, eşik değeri yükseldiğinde hafifçe düşüş eğilimi gösterir (Bkz. Şekil 4.9). Bu eğilim açıkça gösterir ki, ağ topolojisinde tahmin etme maksatlı kullanılabilecek saklı bilgiler vardır. Eşik değeri arttıkça eğitim seti azalacak, bunun sonucunda ortak komşuları olan düğüm çiftlerinin sayısı azalacak ve sonuç olarak ağda yaşanan bu kayıplar sebebiyle bağlantı tahmini zorlaşacaktır. Diğer bir deyişle meydana gelecek bilgi kaybından dolayı mevcut ağda bulunmayan gelecekte oluşması muhtemel yüksek olasılıklı bağlantılar tahmin edilemeyebilir. Bu da performansın düşmesine sebep olur. Uygun eşik değeri seçme problemi büyük bir problemdir. Tez çalışmamızda, Neal Z.’nin [115] önerdiği yöntemden esinlenilerek ağırlıklandırılmış tek parçalı ağda en yüksek frekansa sahip kenar ağırlığı eşik değeri olarak seçilir. Modellenen ağda bu değer 10 olarak tespit edilmiştir. İlk sonuçlar $\alpha = 10$ eşik değerine göre hesaplanmıştır. Daha sonra bu eşik değerinin üstüne çıkıldıkça tahmin kalitesinin düştüğü fakat PL kümesi azaldığından dolayı çalışma zamanının kısaldığı, eşik değerinin altına düştükçe tahmin kalitesinin arttığı fakat PL kümesi büyüdüğünden dolayı çalışma zamanının arttığı gözlemlenir. Şekil 4.8(a) ve Şekil 4.9(a)’da görüldüğü gibi $\alpha = 8$ değerinde $n = 1000$ için önerilen algoritmanın AUC kalite ölçütüne göre bağlantı tahmini kalite değeri 0,98, Kesinlik kalite ölçütüne göre ise 1’dir. Tahmin kalitesi $\alpha = 8$ eşik değerinde yüksek değerlere sahip olduğundan yöntem daha alt değerli eşikler için test edilmemiştir.



Şekil 4.10. Algoritmaların farklı eşik değerlerindeki bağlantı tahmini kalitelerinin kesinlik-anma eğrileriyle kıyaslanması

Şekil 4.10, her bir algoritmanın dört ayrı eşik değerine göre önerilen bağlantı tahmini yöntemiyle elde edilen potansiyel bağlantı kümesindeki kesinlik-anma eğrisi değerlendirme ölçütüyle hesaplanan tahmin performanslarını gösterir. Kesinlik-anma eğrileri *PL* tahmin kümesinin tüm eşik değerlerindeki kesinlik ve anma değerleri arasındaki değişimi gösterir. Şekil 4.10'da gösterildiği gibi, en iyi performanslar $\alpha = 8$ ile güçlendirilmiş ağ versiyonunda önerilen yeni yakınlık ölçütü algoritmasıyla elde edilmiştir. Önerilen algoritma, bağlantı tespitinde tekrarlanabilir bir hassasiyet ve anma sağlar. Eşik değeri arttıkça, güçlendirilmiş ağ üzerindeki bilgi kaybına bağlı olarak her algoritmanın anma performansının azaldığına da dikkat edilmelidir.



5. İKİ PARÇALI AKADEMİK BİLGİ AĞI İÇİN GELİŞTİRİLEN ZAMAN FARKINDALI BAĞLANTI TAHMİNİ YÖNTEMİ

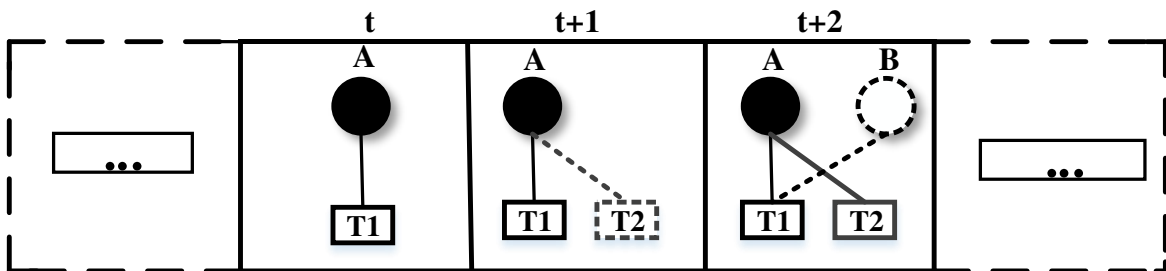
Akademik konu önerisi yazarların kendi uzmanlık alanlarındaki popüler çalışma konularının bulunması sorununa yönelik geliştirilmiş etkili bir çözümdür. Tezin bu bölümünde, bir önceki bölümde anlattığımız iki parçalı akademik bilgi ağının, gelişen yapısını ve ağdaki düğüm çiftleri ile ilgili geçici olayları evrimleştiren yeni bir yakınlık ölçütü algoritması öneriyoruz. Bu amaçla, öncelikle Bölüm 4.1’de bahsedilen veri toplama ve temel bilgi çıkarma teknikleri kullanılarak işlediğimiz veri kümesiyle iki parçalı yazar-konu ağı oluşturulacaktır. Daha sonra önerilen yakınlık ölçütü algoritması için iki parçalı ağların gelişen yapısı anlatılacaktır. Bu bölümde önerilen yöntemde potansiyel bağlantıların bulunması dördüncü bölümde verilen yöntemle bulunur. Fakat, potansiyel bağlantı kümesindeki her bir elemanın benzerlik skoru bu bölümde önerilecek zaman farkındalı yakınlık ölçütü algoritmasıyla hesaplanacaktır. Önerilecek yeni yakınlık ölçütü algoritmasının amacı, iki düğüm arasındaki bir bağlantının gücüyle ilgili öne sürülen Homans'ın fikrini [25], iki düğümün birbirleriyle ne sıklıkla etkileştikleri ve bir önceki bölümde önerilen fikri, iki düğüm arasındaki düğümlerin ortak komşu sayıları, ortak komşularının ağ içerisinde ki önem dereceleri ve iki parçalı ağdaki önem derecelerinin ne kadar büyük olduğuyla ilgilidir, birleştirmektir. Önerilen bu yeni yöntemin en önemli katkısı, bağlantıların kalitesini değerlendirmek amacıyla düğümlerin ve komşularının yakın çevresinde meydana gelen olayların ne zaman oluştuğuna bakılmaksızın ağın mevcut durumunu kullanan yakınlık ölçütleri yerine, iki parçalı ağın evrimsel gelişimini kullanan zaman farkındalı yeni bir bağlantı tahmini algoritmasının önerilmesidir. Ağın evrimsel gelişiminde, düğümler arası etkileşimlerin gücünde artma ya da azalma meydana gelebilir. Önerilen yöntem, meydana gelen bu etkileşimlere dayalı bir yöntemdir. Bildiğimiz kadarıyla, önerilen yöntem iki parçalı ağlarda zaman farkındalığının dikkate alındığı ilk girişimdir. Gerçek veriler üzerinde yapılan deneysel sonuçlar, önerilen yöntemin başarısının umut verici olduğunu gösterir.

5.1. Akademik Bilgi Ağının Oluşturulması

Önerilen yöntemin deneysel sonuçlarını almak amacıyla Bölüm 4.1’de sunulan veri toplama ve ön işleme adımı sonucunda yazarların akademik çalışmalarını içeren iki parçalı yazar-konu ağı oluşturulur. 2000-2016 yılları arasındaki verilerle $G = (U, V, E)$ iki parçalı ağ, 2000-2008 yılları arasındaki verilerle $G_{eğitim} = (U, V, E_{eğitim})$ ve 2009-2016 yılları arasındaki verilerle $G_{test} = (U, V, E_{test})$ iki parçalı ağları modellenir. G iki parçalı ağ 67.677 yazar ve 122.852 konu düğümlerinden ve bu düğümlerin oluşturduğu 1.260.559 bağlantıdan oluşur. Önerilen zaman farkındalı bağlantı tahmini yöntemimizde eğitim ağı evrimsel gelişimiyle birlikte analiz edilecektir. Bunun için önerilen zaman farkındalı iki parçalı ağ yapısı aşağıdaki bölümde detaylı olarak anlatılacaktır.

5.2. Zaman Farkındalı İki Parçalı Ağ Yapısı

Gerçek dünya sistemlerinde bireyler arasındaki ilişkiler sürekli değişmekte ve gelişmekte olduğundan gerçek dünya verilerinden oluşturulmuş iki parçalı sosyal ağlar da sürekli değişmektedir. Bu ağların zamansal gelişimi, düğümler ve bağlantılarla ilgili değerli bilgiler içerdiğinden bağlantı tahmini problemlerinde dikkate alınması gereken potansiyel bir kaynaktır. Bu bölümde zaman farkındalı iki parçalı ağların genel yapısı tanıtılacaktır. Şekil 5.1’de zaman farkındalı iki parçalı ağların evrimsel gelişimine örnek verilmiştir.



Şekil 5.1. Zaman farkındalı iki parçalı ağın evrimsel gelişimi

Zaman farkındalı iki parçalı ağ yapısında, mevcut ağ farklı zaman aralıklarında ağ durumlarını temsil eden önceden belirlenmiş birkaç zaman dilimli anlık görüntü ağlarına ayrılır. Daha sonra ardışık zaman dilimlerindeki anlık görüntü ağları birleştirilerek çerçeveler oluşturulur. Oluşturulan her bir çerçeve boyutu aynıdır. Bu tez çalışmasında, iki parçalı

ağların evrimsel gelişimini temsil eden zamansal ağ modeli tanımlanır. Ağın her bir anlık görüntüsü ağın evrimsel gelişimindeki bir zaman dilimidir. $G = (U, V, E)$, U ve V gibi birbirinden bağımsız düğüm kümelerinden oluşan iki parçalı ağdır. $G_T = (U_T, V_T, E_T)$, G ağının sadece T zamanına kadar gözlemlenen kenarlarını içeren alt ağdır. $[G_1, G_2, \dots, G_T]$, 1'den T zamanına kadar olan çizgelerin birleşiminden oluşan anlık görüntü çerçevesidir. n anlık görüntü çerçeve sayısı ise $F = \{F_1, F_2, \dots, F_n\}$ şeklinde tanımlanır.

Matematiksel olarak ifade etmek gerekirse:

Tanım 9 (Zaman Farkındalı İki Parçalı Ağ): $G_T = (U_T, V_T, E_T)$, zaman farkındalı iki parçalı ağdır. U_T ve V_T iki parçalı G_T ağının iki ayrık düğüm kümesidir ve $E_T \subseteq |U_T| \times |V_T|$ ayrık düğüm kümeleri arasındaki ilişkileri temsil eden bağlantı kümesidir. G_T , zaman farkındalı iki parçalı ağ yapısı T zaman aralığında oluşmuş bağlantıları gösterir. E_T bağlantı kümesinde, her bir $u_i \in U_T$ ve $v_j \in V_T$ düğümleri arasında t zamanında oluşan bağlantı, (u_i, v_j, t) ile temsil edilir. t , T zaman kümesindeki bir zaman etiketidir ($T - m \leq t \leq T$), m G_T ağındaki zaman etiketlerinin sayısıdır. Aynı U_T ya da V_T düğüm kümeleri arasında oluşmuş bağlantı yoktur. Diğer bir deyişle, $(u_i, v_j) \in E_T$ gibi her bir kenar için, $u_i \in U_T, v_j \in V_T$ ya da $v_j \in U_T, u_i \in V_T$ olmalıdır.

G_T zaman farkındalı iki parçalı ağ, sırasıyla $G_{T-m}, \dots, G_T$ grafiklerini temsil eden $A_{T-m}, \dots, A_T$ komşuluk matrislerinin birleştirilmesinden elde edilen $A_{T-m,T}$ komşuluk matrisiyle temsil edilebilir. $A_t(ij)$ binary değeri, t zamanında G_T ağında u_i ve v_j düğümleri arasında bağlantı olup olmadığını gösterir.

$$A_t(ij) = \begin{cases} 1, & \text{if } (u_i, v_j) \in E(G_t), \\ 0, & \text{aksi taktirde.} \end{cases} \quad A_{T-m,T} = \sum_{t=T-m}^T A_t.$$

Eğer G_T iki parçalı ağ ω zaman periyodunda n adet anlık görüntü alt ağına bölünürse, zaman farkındalı iki parçalı ağ yapısı aşağıdaki F ile temsil edilir [28]. G_T ağının bir alt ağı olan G_t , sadece t zamanına kadar oluşmuş bağlantıları içerir. $F = \{F_1, F_2, \dots, F_n\}$, n çerçeveli evrimsel bir süreçte $(n - 1)$ adet A evrimsel geçiş vardır, F ve A evrimsel gelişim yapıları aşağıdaki gibidir:

$$F = \{[G_{T-n\omega+1}, G_{T-n\omega+2}, \dots, G_{T-(n-1)\omega}], \dots,$$

$$[G_{T-2\omega+1}, G_{T-2\omega+2}, \dots, G_{T-\omega}], [G_{T-\omega+1}, G_{T-\omega+2}, \dots, G_T]\}. \quad (5.1)$$

$$A = \{A_1, A_2, \dots, A_{n-1}\},$$

$$A_i = \{F_i \rightarrow F_{i+1}\}. \quad (5.2)$$

5.2.1. Zaman Farkındalı İki Parçalı Akademik Bilgi Ağının Gelişen Yapısı

Zaman farkındalı gelişen yapı, modellenen iki parçalı akademik bilgi ağında yazarların farklı yıllardaki zaman aralıklarına göre çalıştıkları konu etiketleriyle olan bağlantılarına göre ağın durumlarını temsil eder. Örneğin, $\{[2000,2002]\}$ zaman aralığındaki bir çerçeve F_i , tüm yazarların 2000-2002 yılları arasındaki tüm akademik makalelerinden çıkartılan konu etiketleriyle arasında oluşan bütün bağlantıları içeren alt ağıdır.

Önerilen zaman farkındalı yakınlık ölçütü algoritmasını test etmek amacıyla [2000-2008] yılları arasındaki verilerle oluşturulan $G_{eğitim} = (U, V, E_{eğitim})$ ağı evrimsel gelişimiyle birlikte analiz edilir. Bu ağ $G_T = (U_T, V_T, E_T)$, $T = [2000,2008]$ zaman farkındalı iki parçalı ağ yapısıyla temsil edilir. Tanım 9'a göre $\omega = 3$ yıllık zaman periyodunda $n = 3$ adet çerçeveye bölünürse, iki parçalı ağın zaman farkındalı evrimsel yapısı aşağıdaki gibi olur:

$$F = \{[G_{2000}, G_{2001}, G_{2002}], [G_{2003}, G_{2004}, G_{2005}], [G_{2006}, G_{2007}, G_{2008}]\}.$$

Bu yapı, 2000 ve 2008 yılları arasındaki gözlemlenen bağlantıları temsil eden iki parçalı ağın evrimini analiz etmek için kullanılabilir. Yukarıda verilen zaman aralıklarına göre üç çerçeveli bir akademik bilgi ağının gelişen yapısı aşağıdaki gibidir.

$$F = \{F_1, F_2, F_3\},$$

$$A = \{A_1, A_2\}, A_1 = \{F_1 \rightarrow F_2\}, A_2 = \{F_2 \rightarrow F_3\}.$$

Önerilen yakınlık ölçütü algoritması için genel tahmin edici yapısı da aşağıdaki gibidir:

$$F = \{F_1, F_2, F_3 \rightarrow F_4\}:$$

$$F = \{[G_{2000}, G_{2001}, G_{2002}], [G_{2003}, G_{2004}, G_{2005}], [G_{2006}, G_{2007}, G_{2008}] \\ \rightarrow [G_{2009}, \dots, G_{2016}]\}.$$

Burada F_4 , tahmin edici çerçevedir. Diğer bir deyişle test için kullanılan 2009-2016 yılları arasındaki verilerle oluşturulan G_{test} ağıdır.

5.3. İki Parçalı Ağlarda Zaman Farkındalı Bağlantı Tahmini Yöntemi

Önceki bölümde, iki taraflı ağlarda gelecekte oluşması muhtemel potansiyel bağlantıları bulmak için yeni bir bağlantı tahmini yöntemi ve ağın evrimsel gelişiminin göz ardı edildiği yeni bir yakınlık ölçütü algoritması sunduk. Literatürde yapılan çalışmalar analiz edildiğinde, ağın evrimsel gelişiminde meydana gelen olayların dikkate alınmasının bağlantı tahmin performansını ciddi oranda etkilediği fark edilmiştir. Bu bölümde zaman kavramının dikkate alındığı ağın evrimsel gelişimiyle birlikte analiz edildiği yeni bir yakınlık ölçütü algoritması önerilecektir. Önerilecek zaman farkındalı yakınlık ölçütü algoritmasını test etmek için Bölüm 4.1’de anlatılan veri toplama ve temel bilgi çıkarımı bölümünden edilen veriler kullanılarak iki parçalı bir ağ oluşturulur. İki parçalı ağdan gelecekte oluşması muhtemel potansiyel bağlantıların (PL) çıkartılması önceki bölümde sunulan yöntemle aynıdır. Önerilecek zaman farkındalı yakınlık ölçütü algoritmasıyla potansiyel bağlantı kümesindeki her bir potansiyel bağlantının oluşma olasılığı hesaplanacaktır.

5.3.1. Zaman Farkındalı Yakınlık Ölçütü Algoritması

Önerilen bağlantı tahmini yönteminde, PL kümesinin her bir elemanına tahmin puanı atanır, bu da her bir düğüm çifti arasındaki bağlantının oluşma olasılığının bir ölçümüdür. PL kümesinin bir ögesinin tahmin olasılığını hesaplamak için, kapsadığı örüntülerin evrimsel gelişimini ele alan, zamana duyarlı bir yakınlık ölçütü algoritması önerilir. Algoritmada ilk olarak potansiyel bağlantı kümesinin çıkartıldığı $G_{eğitim}$ ağı, Bölüm 5.2’de ayrıntıları verilen zaman farkındalı ağ yapısına dönüştürülüp evrimsel gelişiminin incelenmesi amacıyla Tanım 9’da verilen zaman farkındalı çerçevelere bölünür. İkinci olarak, her bir çerçeve için PL kümesinin elemanı tarafından kapsanan örüntü ağırlığı hesaplanır. Hesaplanan örüntü ağırlığı, çerçeveler arası geçişlerde kullanılan kenar ağırlığı

olarak dikkate alınır. Son olarak, PL kümesine ait bir bağlantının final skoru, önerilen olay tabanlı yakınlık ölçütü algoritmasıyla hesaplanır. Önerilen algoritma bu bölümde detaylı olarak anlatılacaktır.

5.3.2. İki Parçalı Ağlarda Zamansal Olaylar

İki parçalı ağlarda farklı kümelerle ait düğüm çiftleri arasında zaman içerisinde yeni bağlantılar oluşabilir veya var olan bağlantılar yok olabilir. Düğüm çiftlerini bir durumdan (bağlantılı veya bağlantılı olmayan) diğerine taşıyan bu hareket ağ içerisinde gerçekleşen zamansal bir olay olarak tanımlanır. Önceden belirlenmiş bir yakınlık ölçütü algoritmasına göre ağırlıklandırılmış iki parçalı bir ağda, bir çerçeveden diğerine geçerken iki düğüm arasındaki bir bağlantının ağırlığı artıp azalabilir. Meydana gelen bu olaylar tutarlı, güçlendiren ve zayıflatan durum olmak üzere birbirini takip eden üç türden biri olarak kategorize edilebilir.

(1) Tutarlı Olay

İki parçalı bir ağda tutarlı durum türü, $(\omega - 1)$ 'inci çerçevede bir düğüm çifti için hesaplanan bağlantı ağırlığı ω 'ninci çerçevede, önceden belirlenen bir durum değişim oranından daha fazla yükselmiyor veya daha fazla azalmıyorsa yani bir çerçeveden diğerine geçerken düğüm çiftleri arasındaki bağlantının gücü korunuyorsa tutarlı olay meydana gelir. $(\omega - 1)$ 'inci çerçeveden ω 'ninci çerçeveye geçiş boyunca bir tutarlı olayı ele alırken iki parçalı ağdan yansıtılmış tek parçalı ağdaki her bir (A, B) düğümü için tutarlı durum skoru $C_\omega(A, B)$ 'yi matematiksel olarak tanımlamak gerekirse:

$$C_\omega(A, B) = \begin{cases} c & \text{eğer } W_{\omega-1}(A, B) - r \cdot W_{\omega-1}(A, B) < W_\omega(A, B) < W_{\omega-1}(A, B) + r \cdot W_{\omega-1}(A, B) \\ 0 & \text{aksi taktırde.} \end{cases} \quad (5.3)$$

Genelliği kaybetmeksizin burada $W_{\omega-1}(A, B)$ ve $W_\omega(A, B)$, sırasıyla $F_{\omega-1}$ ve F_ω çerçevelerindeki A ve B düğümleri arasındaki bağlantının önemini temsil eder. r değeri, çerçeveler arasındaki durum değişimleri bulunurken kullanılan önceden belirlenmiş sabit bir orandır ($0 < r \leq 1$). c sabit değeri ise iki parçalı ağda çerçeveler arası geçişlerde ilgili

düğüm çifti arasındaki bağlantının korunduğu tutarlı durumlar pozitif bir skordur ($0 < c \leq 1$).

(2) Güçlendiren Olay

İki parçalı bir ağda güçlendiren durum türü, $(\omega - 1)$ 'inci çerçevede bir düğüm çifti arasında hesaplanan ilişki ağırlığı ω 'ninci çerçevede, önceden belirlenen bir durum değişim oranından daha fazla yükseliyorsa yani bir çerçeveden diğerine geçerken düğüm çiftleri arasındaki bağlantının gücü artıyorsa güçlendiren olay meydana gelir. $(\omega - 1)$ 'inci çerçeveden ω 'ninci çerçeveye geçiş boyunca güçlendiren türde bir olayı ele alırken iki parçalı ağdan yansıtılmış tek parçalı ağdaki her bir (A, B) düğümü için güçlendiren durum skoru $S_\omega(A, B)$ 'yi matematiksel olarak tanımlamak gerekirse:

$$S_\omega(A, B) = \begin{cases} s & \text{eğer } W_\omega(A, B) > W_{\omega-1}(A, B) + r \cdot W_{\omega-1}(A, B) \\ 0 & \text{aksi taktirde.} \end{cases} \quad (5.4)$$

s sabit değeri, bir çerçeveden diğerine geçişte ilgili düğüm çifti arasındaki bağlantı güçlendiğinden dolayı atanan pozitif bir ödül değeridir ($0 < s \leq 1$).

(3) Zayıflatan Olay

Zayıflatan olaylar güçlendiren olayların tam tersidir. Bu olay türü, bir düğüm çifti arasındaki bağlantının ağırlığı $(\omega - 1)$ 'inci çerçeveden ω 'ninci çerçeveye geçişte önceden belirlenen durum değişim oranından daha fazla azalıyorsa meydana gelir. $(\omega - 1)$ 'inci çerçeveden ω 'ninci çerçeveye geçiş boyunca zayıflatan türde bir olayı ele alırken iki parçalı ağdan yansıtılmış tek parçalı ağdaki her bir (A, B) düğümü için zayıflatan durum skoru $D_\omega(A, B)$ 'yi matematiksel olarak tanımlamak gerekirse:

$$D_\omega(A, B) = \begin{cases} d & \text{eğer } W_\omega(A, B) < W_{\omega-1}(A, B) - r \cdot W_{\omega-1}(A, B) \\ 0 & \text{aksi taktirde.} \end{cases} \quad (5.5)$$

d sabit değeri, bir çerçeveden diğerine geçişte ilgili düğüm çifti arasındaki bağlantı zayıfladığından dolayı atanan negatif bir ceza değeridir ($-1 \leq d < 1$).

İki parçalı ağlarda zaman farkındalı bağlantı tahmini yapılırken ağdaki her bir düğüm çiftinin evrimsel gelişimi; tutarlı, güçlendiren ve zayıflatan olaylarla incelenir. Her bir olay türü için belirlenen ödül ve ceza puanları kullanılarak bağlantı ağırlığı hesaplanır.

Tutarlı, güçlendirici ve zayıflatan olaylar için verilen c , s ve d değerleri, geçerlilik kümesindeki bağlantı tahmin performansının değerlendirilmesiyle deneysel olarak belirlenebilir. Geçerlilik kümesi, tahmin çerçevesinden önceki son çerçevedir.

5.3.3. İki Parçalı Ağlarda Zaman Farkındalı Yakınlık Ölçütü Algoritması

İki parçalı ağdan çıkartılan potansiyel bağlantı kümesinde ki her bir düğüm çiftinin kalitesini ölçmek amacıyla daha önce yapılmamış yeni bir zaman farkındalı yakınlık ölçütü algoritması öneriyoruz. Bu maksatla, öncelikle potansiyel bağlantı kümesindeki her bir düğüm çiftinin kapsadığı örüntü elemanlarının her bir çerçevedeki ağırlığı hesaplanır ve hesaplanan örüntü ağırlığı çerçeveler arası durum geçişlerinde kullanılan kenar ağırlığı olarak hesaba katılır. Örneğin, ω 'ninci çerçevede A ve B düğümlerinin örüntü ağırlığı $W_\omega(A, B)$ 'dir. PL 'nin kapsadığı örüntüler ne kadar önemliyse kapsanan örüntüde ki her bir bağlantının ağ içerisinde ki ağırlığı da bir o kadar önemlidir. Ağlar dinamik bir yapıya sahip olduğundan örüntü ağırlığının evrimsel gelişimi de çok önemlidir. $\{A, B\}$ örüntüsünün zaman farkındalı ağırlığını hesaplarken her evrimsel süreç noktasında diğer bir deyişle her bir çerçevede üç temel faktör dikkate alınır:

- (1) A ve B düğümlerinin iki parçalı ağdaki ortak komşu sayıları
- (2) A ve B düğümlerinin iki parçalı ağdaki ortak komşularının dereceleri
- (3) A ve B düğümlerinin iki parçalı ağdaki dereceleri

Zaman Farkındalı Örüntü Ağırlığı: $G_{eğitim} = (U, V, E_{eğitim})$, eğitim ağının her bir çerçevesi için zamansal ağlara ayrıştırılır. Zamansal ağlar aşağıdaki gibidir:

$$\{G_{f1} = (U, V, E_{f1})\}, \{G_{f2} = (U, V, E_{f2})\}, \dots, \{G_{fs} = (U, V, E_{fs})\},$$

k_A, k_B ve k_z parametreleri n -çerçeveveli evrimsel süreçte $G_{f1}, G_{f2}, G_{f3}, \dots, G_{fn}$ iki parçalı anlık görüntü ağlarında, sırasıyla A, B ve z düğümlerinin dereceleridir. $W(A, B)$ kenar ağırlığı aşağıda verilen denklem 5.5 ile hesaplanır. $\Gamma(A), \Gamma(B)$ sırasıyla A ve B düğümlerinin iki parçalı anlık görüntü ağlarındaki komşularıdır.

(A, B) kenarının ağırlığı, her bir anlık görüntü karesinde aşağıdaki algoritmayla hesaplanır:

$$W(A, B) = \frac{2}{k(A)+k(B)} \sum_{z \in \Gamma(A) \cap \Gamma(B)} \frac{1}{k(z)} \quad (5.6)$$

$$Skor(A, B) = \sum_{\omega=2}^n \beta(\omega) \cdot [P(A, B, \omega) + \alpha \cdot S(A, B, \omega)], \quad \beta(\omega) = \log(\omega) \quad (5.7)$$

$$P(A, B, \omega) = \begin{cases} C_{\omega}(A, B), & \text{Eğer tutarlı durum oluşursa} \\ S_{\omega}(A, B), & \text{Eğer güçlendiren durum oluşursa} \\ D_{\omega}, & \text{Eğer zayıflatılan durum oluşursa} \end{cases} \quad (5.8)$$

$$S(A, B, \omega) = \sum_{z \in \Gamma(A) \cap \Gamma(B)} P(A, z, \omega) + P(z, B, \omega) \quad (5.9)$$

Önerilen yakınlık ölçütü algoritmasında düğümler etrafındaki bağlantılarla ilgili bilgiler bizim için önemli olduğundan komşuluk tabanlı ölçüt kullanılır. Düğümlerin birbirleriyle olan evrimsel gelişimleri ve ortak komşularıyla olan evrimsel gelişimleri olmak üzere sırasıyla birincil ve ikincil kategori olarak adlandırılan iki kategoride incelenir. $P(A, B, \omega)$, $\omega - 1$ çerçevesinden ω çerçevesine geçişte gözlemlenen (A, B) düğüm çifti için gelişen durum skorunu hesaplar ve bu skor birincil durum skoru olarak tanımlanır. Bu durumda gözlemlenen olay; tutarlı, güçlendiren ve zayıflatılmış olaylardan biri olabilir. $S(A, B, \omega)$, $\omega - 1$ çerçevesinden ω çerçevesine geçişte (A, B) düğüm çiftinin ortak komşularıyla olan gelişen durum skorunu hesaplar ve bu skor ikincil durum skoru olarak tanımlanır. Burada $\Gamma(A)$ ve $\Gamma(B)$, her bir çerçevede sırasıyla A ve B düğümünün komşu dizi setidir. α parametresi, ikincil durumların A ve B arasındaki ilişkiyi ne kadar güçlü etkilediğini gösteren önceden belirlenmiş bir kontrol faktörüdür. Denklem 5.6, her iki kategorideki evrimsel gelişim skorunun toplamı (A, B) düğüm çiftinin zamansal yakınlık ölçütüdür. Burada, $\beta(\omega)$ ile eski olanlara nazaran daha sonra meydana gelen olaylara ağırlık verilmiştir. Bu sayede güncel değişimlerin önemi artırılmıştır. Her bir çerçeveler arası geçişte elde edilen bu

skorların birleştirilmesi düğüm çiftinin özelliklerini açıklayan tahmin edici nitelik olarak kullanılır.

Bir PL tarafından kapsanan herbir örüntü elemanının oluşma olasılığı kapsanan örüntülerin toplam ağırlığına eşittir. $G_{eğitim} = (U, V, E_{eğitim})$ eğitim ağından elde edilen ileride oluşması öngörülen (A, p_i) potansiyel bağlantısının zaman farkındalı final skoru aşağıdaki gibidir:

$$S(A, p_i) = \sum_{\{A, B\} \in \Gamma(A, p_i)} Skor(A, B) \quad (5.10)$$

Denklemden de anlaşıldığı gibi eğer (A, p_i) potansiyel bağlantısı, ne kadar yüksek kalitede durum geçişlerine sahip örüntü kapsıyorsa, o kadar yüksek bağlantı tahmini olasılığına sahiptir. Bu yüzden bir potansiyel bağlantının kapsadığı örüntü elemanının her bir geçiş durumunda ki zamansal yakınlık ölçütü değerlerinin toplamı o potansiyel bağlantının bağlantı tahmini final puanıdır.

5.4. Önerilen Yöntemin Sözde Kodu

Giriş: $G(U, V, E)$ iki parçalı ağ, $G_{eğitim}(U, V, E_{eğitim})$ iki parçalı eğitim ağı, α: eşik değeri; T: $G_{eğitim}$ ağının son zaman etiketi, ω: herbir çerçevenin zaman periyodu, n: çerçeve sayısı, $G_{eğitim}$ ağının gelişen yapısı; $F = \{F_1, F_2, \dots, F_n\}$ $= \{[G_{T-n\omega+1}, G_{T-n\omega+2}, \dots, G_{T-(n-1)\omega}], \dots, [G_{T-2\omega+1}, G_{T-2\omega+2}, \dots, G_{T-\omega}], [G_{T-\omega+1}, G_{T-\omega+2}, \dots, G_T]\}.$ Çıkış: (PL) potansiyel bağlantı kümesi ve bu kümedeki her bir elemanın bağlantı tahmini final skoru.	
Part 1: Tanım 3'e göre G_U yansıtılmış ağırlıklı ağ oluşturulması 1. for iki parçalı ağdaki $\forall A \in U$ do 2. for $\forall x \in \Gamma(A)$ do 3. for $\forall B \in \Gamma(x)$ do 4. $W: (A, B) \rightarrow \Gamma(A) \cap \Gamma(B)$ 5. $E_U = E_U \cup \{(A, B)\}$ 6. end for 7. end for 8. end for	Part 2: Tanım 5'e göre G_U yansıtılmış ağın G_U^α güçlendirilmiş tek parçalı ağına dönüştürülmesi 9. for $\forall A \in G_U$ do 10. for $\forall B \in \Gamma_U(A)$ do 11. if $W(A, B) > \alpha$ then 12. $E_U^\alpha = E_U^\alpha \cup \{(A, B)\}$ 13. end if 14. end for 15. end for
Part 3: Tanım 6'ya göre PL (Potansiyel Bağlantılar) kümesinin çıkartılması 16. for G_U^α güçlendirilmiş tek parçalı ağdaki $\forall A \in U$ do 17. PL_A boş dizi 18. for $\forall K \in \Gamma_U^\alpha(A)$ do 19. for $\forall p \in \Gamma(K)$ do 20. if $(A, p) \notin E_{eğitim}$ then 21. $PL_A = PL_A \cup (A, p)$ 22. end if 23. end for 24. end for 25. $PL = PL \cup PL_A$ 26. end for	Part 4: PL kümesinin her bir elemanının bağlantı tahmini final skorunun zaman farkındalı yakınlık ölçütü algoritmasıyla hesaplanması 27. for $\forall (A, p) \in PL$ do 28. for $\forall B \left(B \in (\Gamma_U(A))_{eğitim} \cap \Gamma(p) \right)$ do 29. for $\forall F_i \in \{F_1, F_2, \dots, F_n\}$ do 30. (A, p) potansiyel bağlantı için Denklem (5.5)'e göre $W(A, C)$ kapsanan örüntü ağırlığını hesapla 31. for herbir F_i için $\forall z \in \Gamma(A)$ do 32. for herbir F_i için $\forall B \in \Gamma(z)$ do 33. Denklem (5.5)'e göre $W(A, z)$ ve $W(z, B)$ hesapla 34. end for 35. end for 36. end for 37. Denklem (5.7), (5.8) ve (5.9)'i kullanarak $G_{eğitim}$ ağının A_i geçiş skorunu hesapla 38. $S(A, p_i) = \sum_{\{A, B\} \in \Gamma(A, p_i)} Score(A, B)$, Denklem (5.10) 39. end for 40. end for

5.5. Uygulama Sonuçları

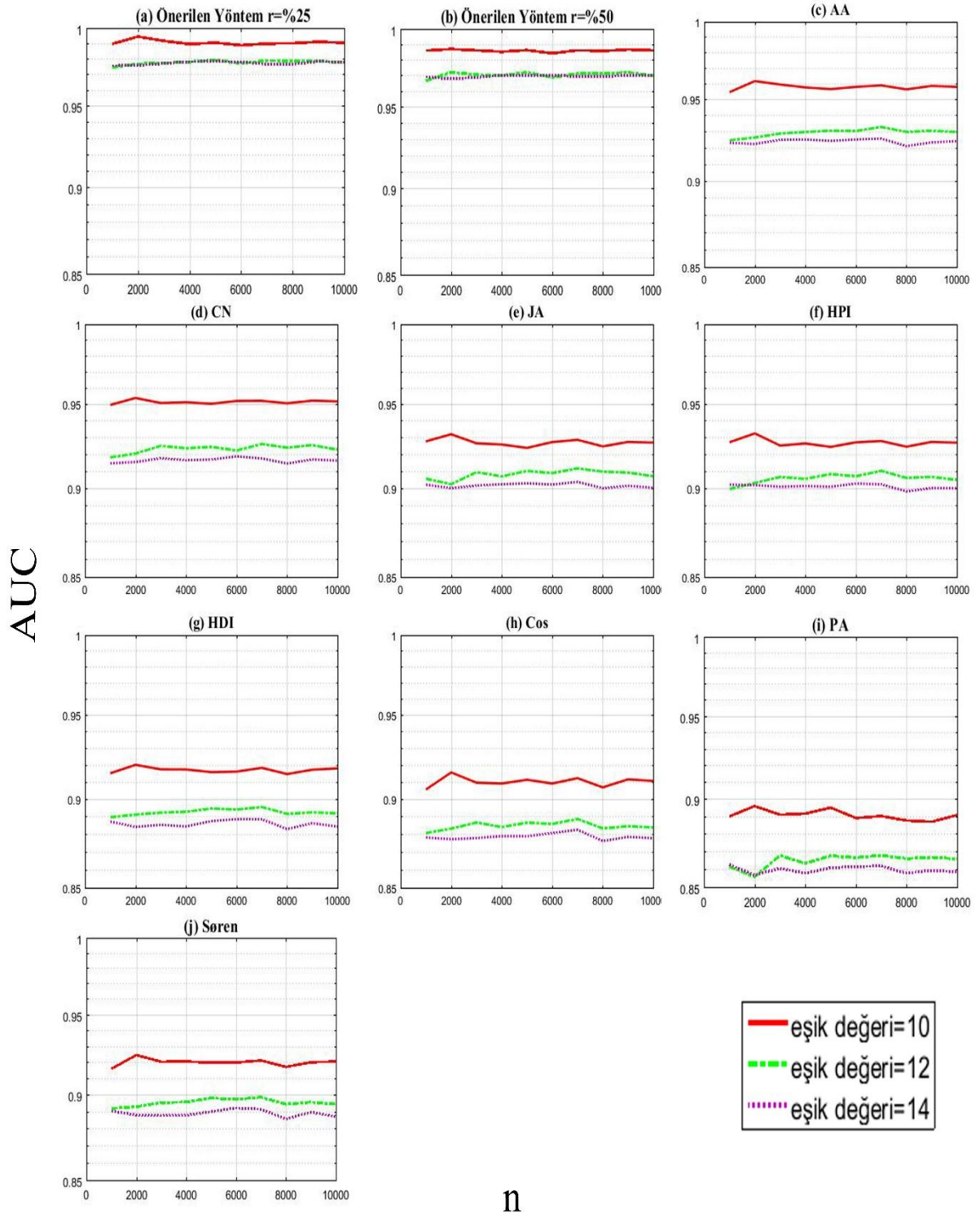
Bu bölümde, önerilen yöntemin uygulanabilirliğini test etmek ve değerlendirmek amacıyla gerçek dünya akademik bilgi verileri kullanarak bazı deneysel uygulamalar gerçekleştirilmiştir. Bu doğrultuda, Bölüm 4.1’de sunulan veri toplama ve önışleme adımı sonucunda yazarların akademik çalışmalarını içeren iki parçalı yazar-konu ağı oluşturulur. [2000,2016] yılları arasındaki etkileşimlerden oluşturulmuş $G = (U, V, E)$ iki parçalı ağ, G_{egitim} ([2000,2008]) ve G_{test} ([2009,2016]) olmak üzere iki adet iki parçalı ağa ayrıştırılır.

Önerilen bağlantı tahmini yöntemi Bölüm 4’te açıklandığı üzere iki parçalı ağlardan yansıtılmış tek parçalı ağların güçlendirilmesine dayalı bir yöntemdir. Bu bölümde önerilen yöntem, $\alpha = 10, \alpha = 12$ ve $\alpha = 14$ eşik değerleri kullanılarak iki parçalı ağın üç ayrı versiyonundan elde edilen potansiyel bağlantılar (PL) için ayrı ayrı test edildi.

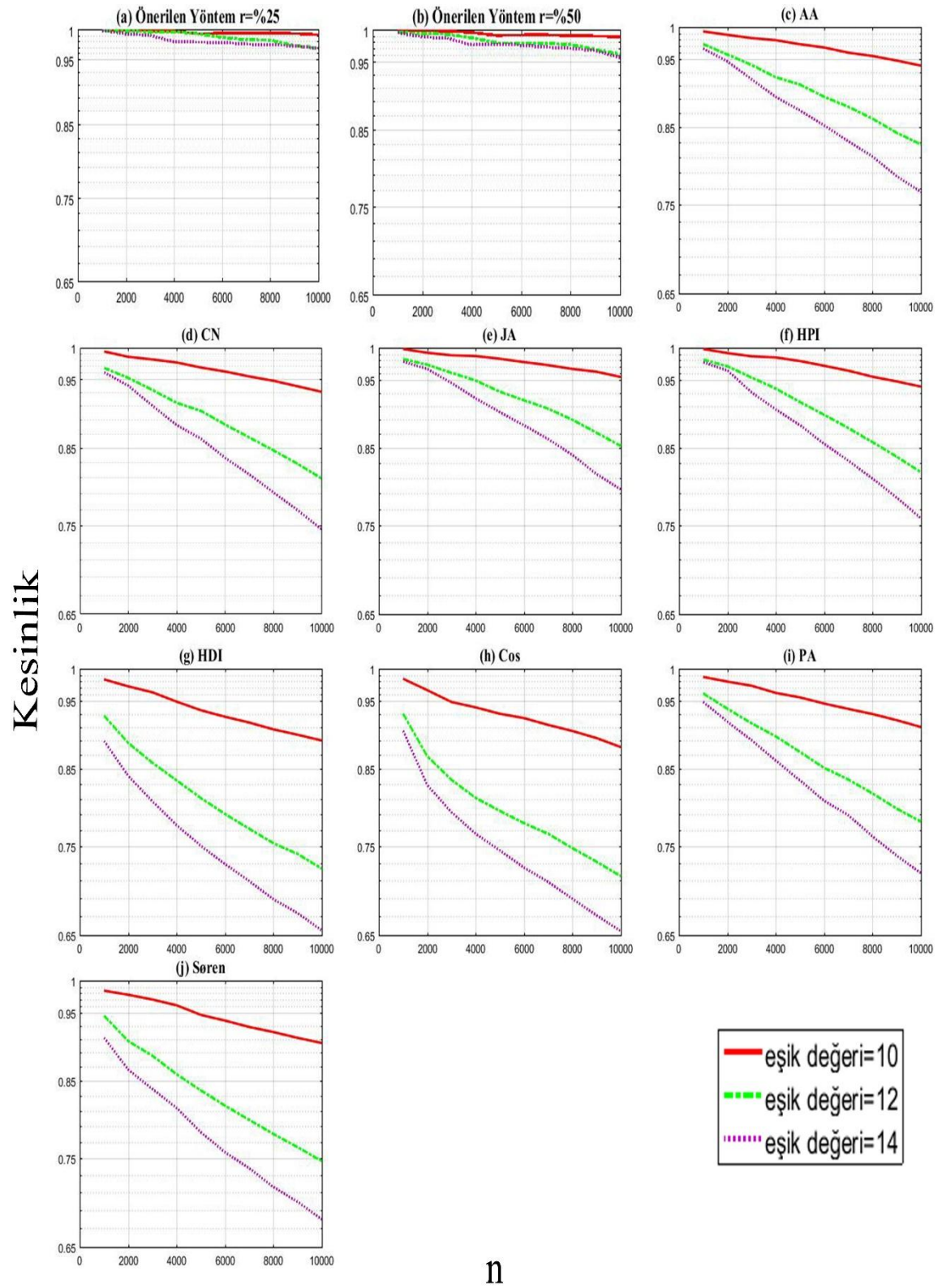
Önerilen algoritmada zaman farkındalı bağlantı tahmini için c, s, d ve α parametrelerinin en uygun değerlerini deneysel olarak değerlendirmek ve belirlemek için kaba kuvvet yöntemi kullanılmıştır. Yapılan testlerle en uygun değerler olarak $c = 0,45, s = 0,85, d = -0,35$ ve $\alpha = 0,07$ bulunmuştur. Yapılan testlerde en iyi performans değerlerinin $c < s$ olduğu durumda elde edildiği görülmüştür. Önerdiğimiz zaman farkındalı yakınlık ölçütü algoritmamız $r = \%25$ durum değişim oranına göre ve $r = \%50$ durum değişim oranına göre iki farklı şekilde hesaplanır. Önerilen zaman farkındalı yakın ölçütü algoritması, olay tabanlı bir yöntemdir. Bu nedenle ağ içerisinde meydana gelen zamansal olayların değişim oranı çok önemlidir. Bu iki farklı durum değişim oranını kullanmanın amacı, çerçeve boyutunun sabit olduğu bir ağda farklı durum değişim oranlarının tahmin sonuçları üzerinde yarattığı etkiyi incelemektir.

Seçilen eşik değerleri kullanılarak üç ayrı potansiyel bağlantı tahmin kümesi elde edildikten sonra kümenin her bir elemanı için önerilen yöntemin dışında literatürde bilinen, ağın evrimsel gelişimini hesaba katmayan $AA, CN, JA, HPI, HDI, Cos, PA, Sprensen$ benzerlik ölçütleriyle karşılaştırıldı. Karşılaştırma için kullanılan algoritmaların detaylı açıklaması Bölüm 3.2’de verilmiştir. Algoritmaların tahmin kalitesi AUC, kesinlik ve kesinlik-anma değerlendirme ölçütleri kullanılarak karşılaştırıldı.

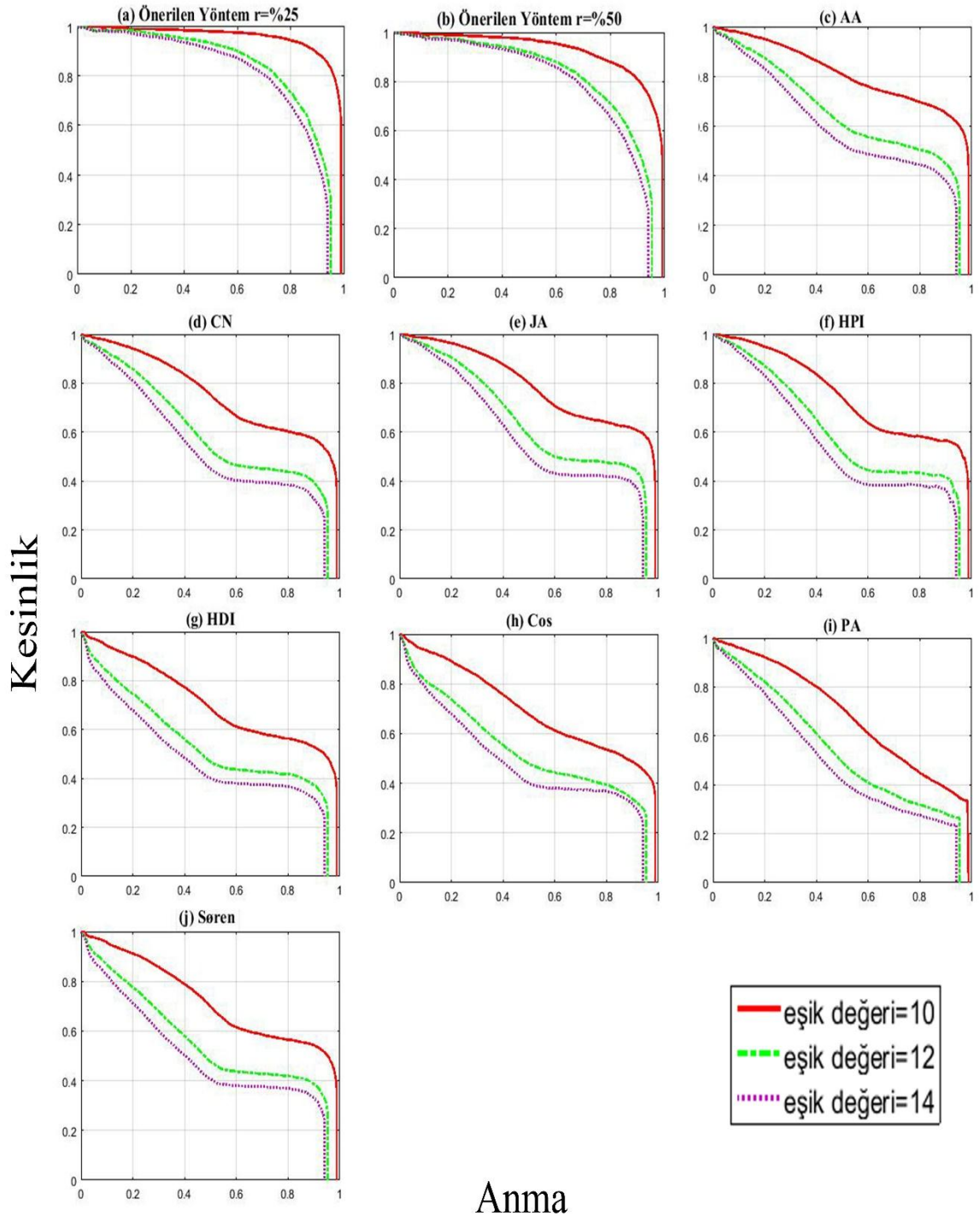
Herbir algoritma için sırasıyla Şekil 5.2’de AUC, Şekil 5.3’te kesinlik ve Şekil 5.4’te kesinlik-anma bağlantı tahmini performansları rapor edilmiştir.



Şekil 5.2. Algoritmaların farklı eşik değerlerindeki bağlantı tahmini kalitelerinin AUC değerlendirme ölçütüne göre kıyaslanması



Şekil 5.3. Algoritmaların farklı eşik değerlerindeki bağlantı tahmini kalitelerinin kesinlik değerlendirme ölçütüne göre kıyaslanması



Şekil 5.4. Algoritmaların farklı eşik değerlerindeki bağlantı tahmini kalitelerinin kesinlik-anma eğrileri değerlendirme ölçütüne göre kıyaslanması

Şekil 5.2, Şekil 5.3 ve Şekil 5.4'te üç farklı güçlendirilmiş tek parçalı ağ için on farklı yakınlık ölçütü algoritmalarının sırasıyla AUC, kesinlik, kesinlik-anma değerlendirme ölçütlerine göre elde edilen bağlantı tahmini performansları rapor edilmiştir. Bu üç şekil incelendiğinde görülmüştür ki, kullanılan değerlendirme ölçütlerinde tüm eşik ve n değerleri için en iyi tahmin performansı $r = \%25$ durum değişim oranıyla önerilen yeni zaman farkındalı yakınlık ölçütü algoritmamız tarafından başarılmıştır. İkinci sırada, önerilen algoritmanın $r = \%50$ durum değişim oranlı versiyonu gelmektedir. Bu durumun en önemli sebebi iki parçalı ağın evrimsel gelişiminin 3 gibi küçük boyutlu çerçeve boyutuyla analiz edilmesidir. Çerçeve boyutunun küçük olduğu yerlerde düşük durum değişim oranıyla fazla sayıda durum değişikliği gözlenirken, yüksek durum değişim oranıyla daha az sayıda durum değişikliği gözlenir. Küçük boyutlu çerçevelerde yüksek durum değişim oranı kullanıldığında ağın evrimsel gelişiminde ki olaylarda bilgi kaybı yaşandığından dolayı bu durum daha düşük performanslı sonuçlar alınmasına sebep olur. Şekillerdeki $r = \%25$ ve $r = \%50$ durum değişim oranlı versiyonların grafikleri karşılaştırıldığında açıkça görülür ki durum değişim oranının tahmin performansına doğrudan etkisi vardır.

Şekil 5.2, Şekil 5.3 ve Şekil 5.4'te gösterildiği gibi, deneysel sonuçlarda kullanılan tüm algoritmaların her bir n değeri için en iyi tahmin performanslarının ağın $\alpha = 10$ ile güçlendirilmiş versiyonunda elde edilirken en kötü tahmin performanslarının ise ağın $\alpha = 14$ ile güçlendirilmiş versiyonundan elde edilmiştir. Bunun en önemli sebebi, güçlendirilmiş ağ oluşturulurken uygun eşik değeri için $\alpha = 10$ değerinin tespit edilmesidir. Bu değerden uzaklaştığımız eşik değerleri için bağlantı tahmini kalitesi düşmektedir. Eşik değeri arttıkça, akademik geçmişleri birbirine daha çok benzeyen yazarlar arasındaki ilişkiler analiz edilmiş ve bu analizler neticesinde konu önerileri yapılmıştır. Ancak, yukarıdaki şekillerde de görüldüğü gibi, bu durum aslında olumsuz bir etkiye sahiptir. Uygun eşik değerinden daha büyük bir eşik değerinin seçilmesi bilgi kaybına sebep olacaktır. Ağdaki bilgi kaybı nedeniyle ağda bulunmayan yüksek olasılıklı bağlantılar tahmin edilemez. Bu performansta bir azalmaya neden olur. Bununla birlikte, eşik değeri arttıkça, eğitim seti azaldığından dolayı ortak komşuları olan düğüm çiftlerinin sayısında bir azalma meydana gelir. Bunun sonucunda bağlantı tahmini zorlaşır. Şekil 5.2, Şekil 5.3 ve Şekil 5.4'ten görülebileceği gibi, tüm algoritmaların bağlantı tahmin nitelikleri, eşik değeri arttığında biraz azalma eğilimindedir. Bu eğilim, ağ topolojisinde tahmin amaçlı kullanılabilecek gizli bilgilerin bulunduğunu açıkça göstermektedir.

Ayrıca, Şekil 5.2, Şekil 5.3 ve Şekil 5.4 (a ve b)'de görüldüğü gibi önerdiğimiz yöntem tüm eşik değerlerinde diğer algoritmalarla kıyasla en iyi sonucu vermiştir. Bunun en önemli sebebi kenar ağırlığı hesaplanırken dikkate alınan parametrelerdir. Örneğin, diğer algoritmalarla örüntü ağırlığı hesaplanırken ağın evrimsel gelişimi göz önüne alınmadan sadece mevcut ağ durumuna bakılarak düğüm derecesi, ortak komşu vs. gibi özelliklerden sadece biri kullanılır. Önerilen yöntemde ise ağın evrimsel gelişiminde hem düğümlerin ortak komşu sayıları, hem ortak komşularının ağ içerisindeki önem dereceleri hem de düğümlerinin iki parçalı ağdaki önem dereceleri göz önüne alındığından ağın topolojik yapısında ki bilgi kaybı minimuma indirilmiştir. Her ne kadar ağın T zamanında topolojik yapısındaki bilgiler korunsa da evrimsel gelişiminde saklı bağlantı tahmininde işe yarayacak çok kıymetli bilgilerin açığa çıkartılması gerekir. Bu büyük boşluğu doldurmak amacıyla önerilen iki parçalı ağları evrimsel gelişimiyle inceleyen zaman farkındalı bağlantı tahmini algoritmamız tüm algoritmalar içerisinde en iyi performansa sahiptir. Aynı zamanda Şekil 5.2, Şekil 5.3 ve Şekil 5.4'te görüldüğü gibi her bir algoritma için en kötü tahmin performansı $\alpha = 14$ eşik değeriyle elde edilen üçüncü versiyon bağlantı tahmini kümesinde (PL) dir.

Bu bölümde edilen sonuçlarda dikkat edilmesi gereken bir diğer hususta, literatürde benimsenen klasik benzerlik ölçütü algoritmalarının üç farklı eşik değerindeki performansları kendi içerisinde kıyaslandığında önerilen yöntemin iki versiyonunda da artan n değerleri için performans değişiminin minimum olmasıdır. Bu durumun en önemli sebebi önerilen yöntemin ağı en iyi temsil edecek özelliklere odaklanmasıdır. Ayrıca, önerilen yöntemin bağlantı tahmini performansının $\alpha = 10$ olarak kabul ettiğimiz uygun eşik değerinden uzaklaştıkça diğer algoritmalarla kıyasla daha kararlı performans göstermektedir. AUC grafiklerinde literatürde benimsenen klasik benzerlik ölçütü algoritmalarının üç farklı eşik değerindeki maksimum ve minimum kalite değerleri arasında büyük bir fark yokken kesinlik grafiklerinde bu fark artmaktadır. Bu durumun temel sebebi, kesinlik ile tahmin kümesinde ki sadece $top - n$ sonuçlarının kalitesi göz önüne alınırken, AUC ve kesinlik-anma değerlendirme ölçütleriyle algoritmanın tüm tahmin kümesi sonuçlarının kalitesinin göz önüne alınmasıdır.

5.6. Farklı Çerçeve Boyutuna Sahip Ağlarda Durum Değişim Oranının Sonuçlara Etkisi

Tezin bu bölümünde farklı çerçeve boyutlarına sahip iki farklı ağda, farklı r durum değişim oranlarının kullanılması durumunda önerilen yöntemin sonuçlarının nasıl değiştiği irdelenmiştir. Bu amaçla, [2000,2016] yılları arasındaki verilerle oluşturulmuş $G = (U, V, E)$ iki parçalı akademik bilgi ağı iki farklı versiyonda eğitim ve test verilerine ayrıştırılarak modellenenecektir.

G iki parçalı ağın ilk versiyonunda; [2000-2013] yıllarındaki verilerle $G_{eğitim} = (U, V, E_{eğitim})$ ve [2014,2015] yılları arasındaki verilerle de $G_{test} = (U, V, E_{test})$ ağı modellenir. G iki parçalı ağın ikinci versiyonunda ise [2000-2011] yıllarındaki verilerle $G_{eğitim} = (U, V, E_{eğitim})$ ve [2012,2015] yılları arasındaki verilerle de $G_{test} = (U, V, E_{test})$ ağı modellenir. Bu bölümde ağın her iki versiyonunda gerçekleştirilecek deneylerde önerilen yöntem $\alpha = 10$ ve $\alpha = 14$ eşik değerleri kullanılarak iki farklı güçlendirici versiyonu için ayrı ayrı test edilmiştir. Bu bölümde ayrıca G_{test} çerçevesinin boyutunu diğer bir deyişle tahmin edici çerçevenin boyutunu $G_{eğitim}$ ağının gelişen yapısındaki her bir çerçeve boyutuyla aynı boyutta çerçevelerden oluşturduk.

- (1) Ağın ilk versiyonunda, zaman farkındalı bağlantı tahmini yöntemimiz için, eğitim ağı Tanım 9'a göre $\omega = 2$ zaman periyodunda $n = 7$ adet anlık görüntü ağına diğer bir deyişle çerçeveye bölünür.

Matematiksel olarak ifade etmek gerekirse,

$$A_1 = \{F_1 \rightarrow F_2\}, A_2 = \{F_2 \rightarrow F_3\}, A_3 = \{F_3 \rightarrow F_4\}, A_4 = \{F_4 \rightarrow F_5\}, A_5 = \{F_5 \rightarrow F_6\}, A_6 = \{F_6 \rightarrow F_7\}$$

Ağın ilk versiyonunda, $F_1, F_2, F_3, F_4, F_5, F_6$ ve F_7 çerçeveleri iki parçalı ağın evrimsel gelişim yapısı olarak belirlenmişken, F_8 ise tahmin edici çerçeve olarak kullanılmıştır.

$$F = \{F_1, F_2, F_3, F_4, F_5, F_6, F_7 \rightarrow F_8\},$$

$$F = \{ [2000,2001], [2002,2003], [2004,2005], [2006,2007], [2008,2009], [2010,2011], [2012,2013] \rightarrow [2014,2015] \}$$

- (2) Ağın ikinci versiyonunda, zaman farkındalı bağlantı tahmini yöntemimiz için, eğitim ağı Tanım 9'a göre $\omega = 4$ zaman periyodunda $n = 3$ adet anlık görüntü ağına diğer bir deyişle çerçeveye bölünür.

Matematiksel olarak ifade etmek gerekirse,

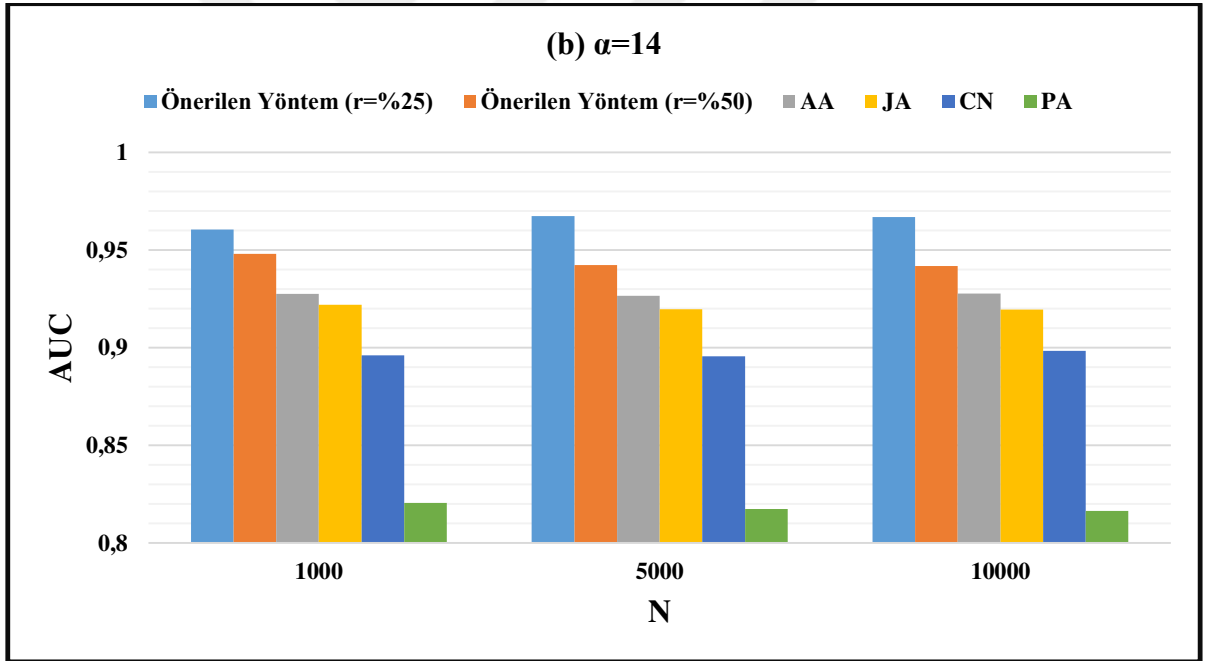
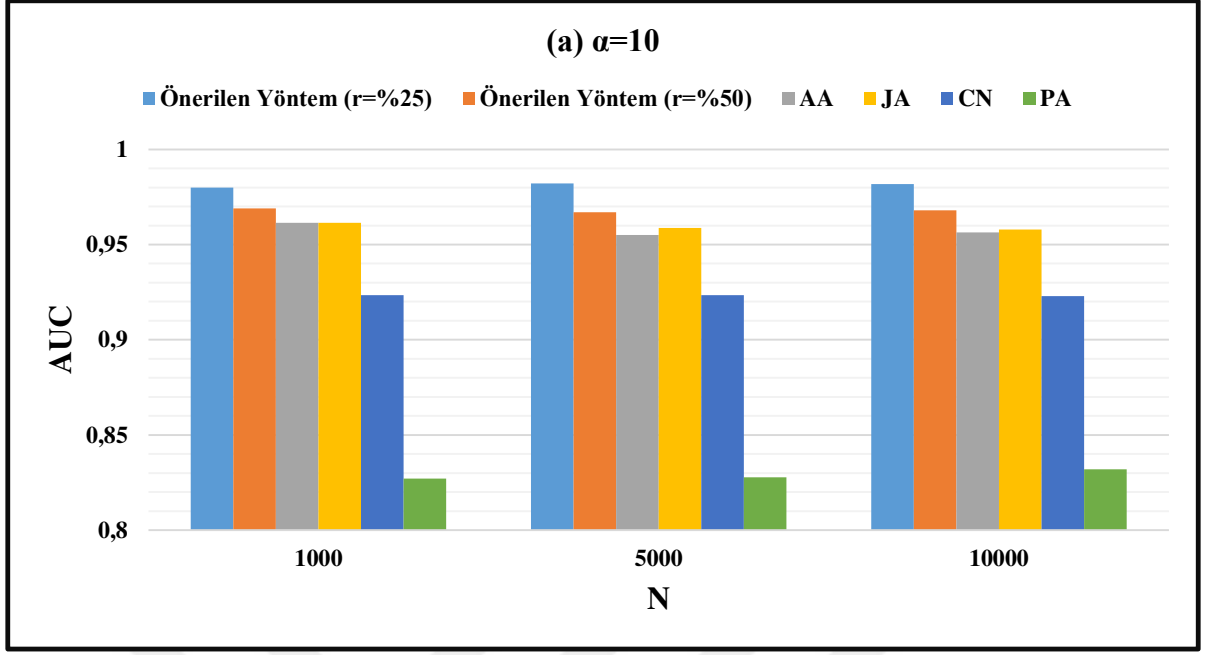
$$A_1 = \{F_1 \rightarrow F_2\}, A_2 = \{F_2 \rightarrow F_3\}$$

Ağın ikinci versiyonunda F_1, F_2 ve F_3 çerçeveleri iki parçalı ağın evrimsel gelişim yapısı olarak belirlenmişken, F_4 ise tahmin edici çerçeve olarak kullanılmıştır.

$$F = \{F_1, F_2, F_3 \rightarrow F_4\},$$

$$F = \{[2000,2003], [2004,2007], [2008,2011] \rightarrow [2012,2015]\}.$$

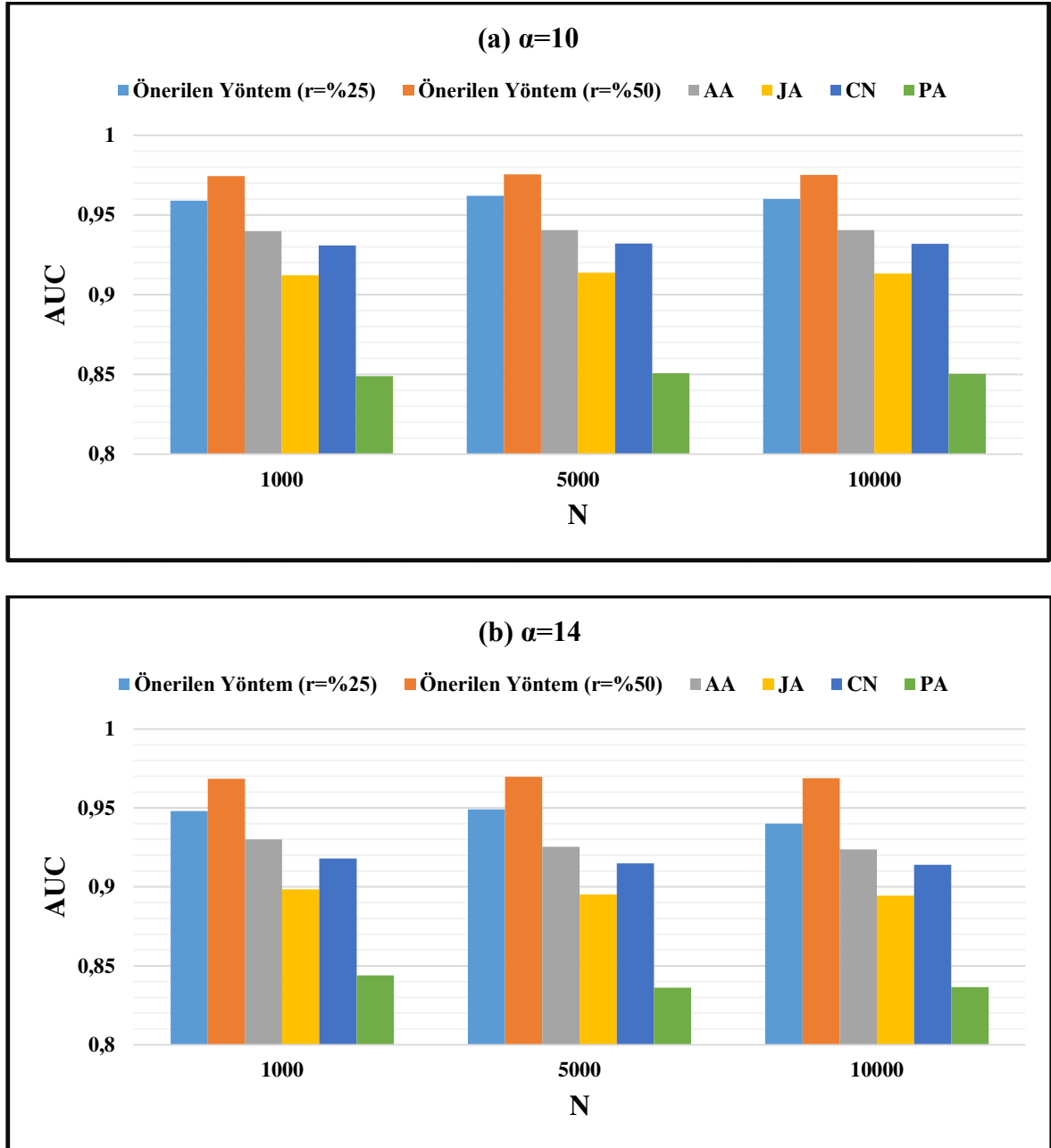
Modellenen her iki ağ versiyonu için bazı deneyler gerçekleştirdik. Deneylerde önerilen yöntemin performansını, Bölüm 3'te detaylı olarak açıkladığımız ağın evrimsel gelişimini dikkate almayan *AA*, *CN*, *JA*, *PA* klasik yakınlık ölçütü algoritmalarıyla karşılaştırdık. Önerilen zaman farkındalı bağlantı tahmininde kullanılmak üzere en uygun c, s, d ve α parametrelerini deneysel olarak belirlemek için geçerlilik kümesi üzerinde kaba kuvvet yöntemi uygulanmıştır. Geçerlilik kümesi tahmin edici çerçeveden önceki son çerçevedir. Yapılan testlerde en uygun parametre değerlerine göre $c = 0,45$, $s = 0,85$, $d = -0,35$ ve $\alpha = 0,07$ olarak tespit edilmiştir. Önerdiğimiz zaman farkındalı yakınlık ölçütü algoritması ilk olarak $r = \%25$ durum değişim oranına göre hesaplanır ve ikincisi $r = \%50$ durum değişim oranına göre hesaplanır. Bu iki benzerlik oranının kullanılmasının amacı, farklı durum değişiklik oranlarının tahmin sonuçlarına olan etkisini incelemektir. Klasik yakınlık ölçütü algoritmalarının performans değerleri ve önerilen zamana duyarlı yaklaşım ölçütü algoritması AUC [32] değerlendirme ölçütü kullanılarak test edilmiştir.



Şekil 5.5. $\alpha = 10$ ve $\alpha = 14$ eşik değeriyle güçlendirilmiş daha küçük aralıklı çerçeveli birinci ağ versiyonunda farklı tahmin edicilerin AUC değerleri

İki parçalı ağı küçük çerçeveli birinci versiyonu için altı yakınlık ölçütü algoritmasının AUC değerlendirme ölçütüne göre performans değerleri Şekil 5.5'te verilmiştir. Çerçeveler arasındaki bağlantıların durum değişiklikleri (tutarlı, güçlendiren ve zayıflatan), r durum değişim oranıyla belirlenir. Bu sebeple 2 gibi küçük boyutlu çerçeveler için düşük oranlı r

durum deęişim oranı deęerleriyle daha fazla sayıda durum deęişikliği gözlemlenirken yüksek oranlı r durum deęişim oranı deęerleriyle daha az sayıda durum deęişikliği gözlenmektedir. Bu durumun tahmin performansı üzerinde doğrudan bir etkisi vardır. Şekil 5.5'te görüldüğü gibi en iyi tahmin performansları tüm eşik deęerlerinde ki n deęerleri için $r = \%25$ durum deęişim oranıyla önerdiğimiz algoritmaya aittir.



Şekil 5.6. $\alpha = 10$ ve $\alpha = 14$ eşik deęeriyle güçlendirilmiş daha büyük aralıklı çerçevesi ikinci ağı versiyonunda farklı tahmin edicilerin AUC deęerleri

Büyük zaman aralıklı çerçeve boyutunda, r durum değiştirme oranının önerdiğimiz zaman farkındalı yakınlık ölçütü algoritmasının performansını nasıl etkilediğini incelemek için farklı bir ağ modelledik. Şekil 5.6 büyük çerçeve boyutlu ağdaki algoritmamızın ve zaman kavramının dikkate alınmadığı diğer dört klasik yakınlık ölçütü algoritmasının sonuçlarını rapor etmiştir. Deneysel sonuçlardan anlaşıldığı üzere büyük çerçeveli ağlarda yüksek durum değişim oranıyla daha iyi performans sonuçları elde edilmiştir. Kısaca özetlemek gerekirse, çerçeve boyutunun küçük olduğu yerlerde düşük durum değişim oranlı ağ daha iyi sonuçlar verirken, çerçeve boyutunun büyük alındığı yerlerde yüksek durum değişim oranlı ağ daha iyi sonuçlar vermektedir.

Şekil 5.5 ve Şekil 5.6'da görüldüğü gibi deneysel sonuçların her iki versiyonunda, algoritmaların farklı eşik değerlerine göre tahmin performanslarını incelediğimizde, her bir algoritma, $\alpha = 10$ değerinde daha yüksek AUC performans skorları elde etmiştir. Bunun nedeni, güçlendirilmiş ağın oluşturulmasında, düğümler arasındaki frekansın en yüksek olduğu kenar ağırlığının, uygun eşik değerine dayanmasıdır. Bununla birlikte, Şekil 5.5 ve Şekil 5.6'da gösterildiği gibi, önerilen yöntem diğer algoritmalara kıyasla tüm eşik değerlerinde en iyi sonucu vermiştir.

6. SONUÇLAR VE ÖNERİLER

Son yıllarda, birçok gerçek dünya sistemi karmaşık ağlar olarak modellenmektedir. E-ticaret ağları (müşteri-ürün: müşterilerin aldıkları ürünlerle bağlandığı), biyolojik ağlar (hastalık-gen, RNA-protein), bibliyografik ağlar (yazar-makale, yazar-atıf) vb. gibi birçok karmaşık ağ doğası gereği iki parçalı yapıya sahiptir. Bu ağlar farklı alanlardan gelmekle birlikte, hepsinin modüler yapısı veya topluluk yapısı benzerdir. İki parçalı ağlar karmaşık ağların modellenmesinde kullanılan özel bir ağ türüdür. Bu ağlar, kenarların yalnızca farklı kümelerdeki düğümler arasında oluştuğu, ancak aynı kümedeki düğümler arasında oluşmadığı, karşılıklı olarak birbirinden bağımsız iki düğüm kümesinden meydana gelir.

Sosyal ağlarda zamanla birçok yeni düğümün veya düğümler arası bağlantıların ortaya çıkması veya var olanların yok olması gibi çeşitli değişiklikler meydana gelmektedir. Sosyal ağlar bu değişikliklerle gelişen dinamik yapılardır. Dolayısıyla bu dinamik yapı, sosyal ağlardaki bağlantıların bir sonraki zaman diliminde ne olacağı sorusunu akla getiriyor.

Bu tez çalışmasında, büyük ölçekli gerçek dünya verilerinden oluşturulmuş iki parçalı akademik bilgi ağında yazarların gelecekte çalışmaya aday oldukları konularla ilgili tavsiyelerde bulunan benzerlik tabanlı yeni bir bağlantı tahmini sistemi geliştirilmiştir. Bu tez çalışmasının literatüre kazandırdığı katkılar aşağıda sıralanmıştır:

- Öncelikle, büyük ölçekli iki parçalı ağ oluşturmak için klasik veri tabanları yerine gerçek dünya sistemlerinden amaca uygun veri toplanmıştır. Toplanan veri ham veridir. Bu verinin işlenmesi için temel bilgi çıkarımı mimarisi önerilmiştir. İşlenen veri kullanılarak iki parçalı akademik bilgi ağı modellenmiştir. Oluşturulan ağ iki parçalı karmaşık bir gerçek dünya ağıdır.
- İki parçalı ağdan, gelecekteki t' zamanda oluşacak potansiyel bağlantıları tahmin etmek amacıyla ağı doğrudan tek parçalı ağlara dönüştüren klasik yöntemlerden farklı olarak ağın omurgası niteliğini taşıyan güçlendirilmiş tek parçalı ağ modeli sunulmuştur. Önerilen yöntem, gelecekte oluşmasını öngördüğü potansiyel bağlantıları sadece güçlendirilmiş ağda varlığını koruyan düğüm çiftleri arasında aramaktadır.

- İki parçalı ağlar tek parçalı ağlara dönüştürüldüğünde bilgi kaybı olmaktadır. Literatürde yapılan çalışmalardan farklı olarak önerilen yöntem hem iki parçalı ağ hem de bu ağdan elde edilen tek parçalı ağ bilgilerini kullanır.
- Gelecekte oluşması öngörülen potansiyel bağlantıların tahmin skorunu hesaplamak için ağın t zamanındaki statik yapısını kullanarak zaman kavramının göz ardı edildiği ve ağın evrimsel gelişimini izleyen zaman farkındalı iki yeni yakınlık ölçütü algoritması geliştirilmiştir. Geliştirilen her iki yakınlık ölçütünde düğümler arası bağlantı ağırlıkları hesaplanırken klasik yaklaşımlardan farklı olarak üç farklı topolojik özellik dikkate alınmıştır.

IEEE Xplorer'dan toplanan 64.677 yazarın [2000,2016] yılları arasındaki akademik çalışmaları kullanılarak oluşturulan iki parçalı ağ üzerinde yapılan uygulamalardan aşağıdaki sonuçlar elde edilmiştir:

- Güçlendirilmiş ağ versiyonları sayesinde elde edilen potansiyel bağlantı tahmini kümesi (PL) küçülmüş ve buna bağlı olarak hesaplama zamanında ciddi azalma sağlanmıştır.
- Güçlendirilmiş ağ için tespit edilmiş uygun eşik değerinden daha yüksek eşik değerlerinin kullanılmasının tahmin performansını olumsuz etkilediği izlenmiştir.
- Önerilen yöntemle hem iki parçalı ağ hem de bu ağdan elde edilen tek parçalı ağ bilgilerinin kullanılması tahmin performansını arttırmıştır.
- Ağdaki tek bir yapısal özelliği dikkate alan klasik yakınlık ölçütleriyle kıyaslandığında önerilen her iki yakınlık ölçütü algoritmasının çok daha iyi sonuçlar verdiği rapor edilmiştir. Yüksek eşik değerli ağ versiyonlarında bile klasik algoritmaların tahmin performansında büyük azalma olurken önerilen yakınlık ölçütü algoritmaları kararlılıklarını bozmamıştır.
- Zaman farkındalı yakınlık ölçütü algoritmasıyla elde edilen deneysel sonuçlar incelendiğinde, çerçeve boyutunun küçük olduğu yerlerde düşük durum değişim oranıyla elde edilen sonuçların daha iyi olduğu, çerçeve boyutunun büyük olduğu yerlerde yüksek durum değişim oranıyla elde edilen sonuçların daha iyi olduğu görülmüştür.

İleriki çalışmalarda, önerilen yöntemin aşağıdaki şekilde geliştirilmesi planlanmaktadır.

- Bu tez çalışmasında modellenen akademik bilgi ağı yazarlar ve çalıştıkları konu etiketlerinden oluşan iki farklı türde düğüm kümesinden oluşur. Sonraki çalışmada yazar, akademik çalışma başlığı ve akademik çalışmanın basıldığı yer (dergi/konferans vb.) düğüm türlerinden oluşan çok modlu ağ yapısı kullanılacaktır. Modellenen bu ağ sayesinde, verilen bir makale başlığı kullanılarak basıma gönderileceği yerlerle ilgili önerilerde bulunulacaktır. Örneğin, makalenin “*sosyal ağ, sosyal ağ analizi, bağlantı tahmini*” konularıyla ilgili olduğu bulunursa önerilen yöntem sayesinde bu makaleye ASONAM ve socialCom gibi doğrudan sosyal ağ analizi ve madenciliği ile ilgili konferanslara gönderilmesi önerilir. Böylelikle bilim insanları hem zamandan kazanç sağlayacak hem de otomatik olarak gerçekleşen analiz sonucu konuya en uygun yerleri bulacaktır.
- Önerilen güçlendirilmiş tek parçalı ağ modeline dayalı bağlantı tahmini yöntemi, farklı alanlardaki gerçek dünya verilerinden oluşturulmuş ağa uygulanacaktır. Bu sayede önerilen yöntemin ulaşım, sağlık, eğitim vb. alanlardaki başarısı da test edilmeye çalışılacaktır.

KAYNAKLAR

- [1] **Wasserman, S., & Faust, K.** (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.
- [2] **Tang, F.** (2017). *Link-Prediction and its Application in Online Social Networks* (Doctoral dissertation, Victoria University).
- [3] **Hwang, S. Y., Wei, C. P., & Liao, Y. F.** (2010). Coauthorship networks and academic literature recommendation. *Electronic Commerce Research and Applications*, 9(4), 323-334.
- [4] **Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M.** (2006, April). Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*.
- [5] **Folino, F., & Pizzuti, C.** (2010). A comorbidity-based recommendation engine for disease prediction.
- [6] **Scellato, S., Noulas, A., & Mascolo, C.** (2011, August). Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1046-1054). ACM.
- [7] **Liben Nowell, D., & Kleinberg, J.** (2007). The link prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7), 1019-1031.
- [8] **Xiang, E. W.** (2008). A survey on link prediction models for social network data. *Science and Technology*.
- [9] **Al Hasan, M., & Zaki, M. J.** (2011). A survey of link prediction in social networks. In *Social network data analytics* (pp. 243-275). Springer, Boston, MA.
- [10] **Katz, L.** (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39-43.
- [11] **Gao, M., Chen, L., Li, B., Li, Y., Liu, W., & Xu, Y. C.** (2017). Projection-based link prediction in a bipartite network. *Information Sciences*, 376, 158-171.
- [12] **Latapy, M., Magnien, C., & Del Vecchio, N.** (2008). Basic notions for the analysis of large two-mode networks. *Social networks*, 30(1), 31-48.
- [13] **Larremore, D. B., Clauset, A., & Jacobs, A. Z.** (2014). Efficiently inferring community structure in bipartite networks. *Physical Review E*, 90(1), 012805.

- [14] **Chang, Y. J., & Kao, H. Y.** (2012, November). Link prediction in a bipartite network using Wikipedia revision information. In *2012 Conference on Technologies and Applications of Artificial Intelligence (TAAI)* (pp. 50-55). IEEE.
- [15] **Nigam, A., & Chawla, N. V.** (2016). Link prediction in a semi-bipartite network for recommendation. In *Asian Conference on Intelligent Information and Database Systems*(pp. 127-135). Springer, Berlin, Heidelberg.
- [16] **Wang, X., Liu, Y., & Xiong, F.** (2016). Improved personalized recommendation based on a similarity network. *Physica A: Statistical Mechanics and its Applications*, 456, 271-280.
- [17] **Sa, H.R. and Pruencio, R.B.C.,** 2011. Supervised link prediction in weighted networks, in: *Proc. of International Joint Conference on Neural Networks*, pp. 2281–2288.
- [18] **Luo, Y., Liu, Q., Wu, W., Li, F., & Bo, X.** (2014, October). Predicting drug side effects based on link prediction in bipartite network. In *Biomedical Engineering and Informatics (BMEI), 2014 7th International Conference on* (pp. 729-733). IEEE.
- [19] **Li, X., & Chen, H.** (2013). Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach. *Decision Support Systems*, 54(2), 880-890.
- [20] **Li, Z., Wang, R. S., Zhang, S., & Zhang, X. S.** (2016). Quantitative function and algorithm for community detection in bipartite networks. *Information Sciences*, 367, 874-889.
- [21] **Kunegis, J., De Luca, E. W., & Albayrak, S.** (2010, June). The link prediction problem in bipartite networks. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems* (pp. 380-389). Springer, Berlin, Heidelberg.
- [22] **Kaya, B.,** 2014. Sosyal ağ modelleme kullanılarak tıbbi veri ağında bağlantı tahmini yöntemleri, *Doktora Tezi*, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ.
- [23] **Xia, S., Dai, B., Lim, E. P., Zhang, Y., & Xing, C.** (2012, August). Link prediction for bipartite social networks: the role of structural holes. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)* (pp. 153-157). IEEE Computer Society.
- [24] **Shams, B., & Haratizadeh, S.** (2016). SibRank: Signed bipartite network analysis for neighbor-based collaborative ranking. *Physica A: Statistical Mechanics and its Applications*, 458, 364-377.

- [25] **Zhang, P., Zeng, A., & Fan, Y.** (2014). Identifying missing and spurious connections via the bi-directional diffusion on bipartite networks. *Physics Letters A*, 378(32-33), 2350-2354.
- [26] **Davis, D., Lichtenwalter, R., & Chawla, N. V.** (2013). Supervised methods for multi-relational link prediction. *Social network analysis and mining*, 3(2), 127-141.
- [27] **Karçı, A., Boy, O.,** (2011). Sosyal Ağların Web Madenciliği ile Analizi ve Ortak Atıf Analizi ile Benzerlik Tahmini, *Elektrik-Elektronik ve Bilgisayar Sempozyumu*,154.
- [28] **Symeonidis, P., Iakovidou, N., Mantas, N., & Manolopoulos, Y.** (2013). From biological to social networks: Link prediction based on multi-way spectral clustering. *Data & Knowledge Engineering*, 87, 226-242.
- [29] **Aslan, S.,** (2014). Yönsüz ağırlıklı hastalık ağında bağlantı tahmini, *Yüksek Lisans Tezi*, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ.
- [30] **Newman, M. E.** (2001). The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2), 404-409.
- [31] **Bollobás, B.** (2001). Random graphs. 2001. *Cambridge Stud. Adv. Math.*
- [32] **Breiman L. ,** (2001). Random forests, *Machine learning*, vol. 45, no. 1, pp. 5-32.
- [33] **Skiena, S.** (1990). Graph isomorphism. *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*, 181-187.
- [34] **Barthélemy, M.** (2011). Spatial networks. *Physics Reports*, 499(1-3), 1-101.
- [35] **Barabasi, A. L., & Oltvai, Z. N.** (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2), 101.
- [36] **Rapoport, A., & Rebhun, L. I.** (1952). On the mathematical theory of rumor spread. *The bulletin of mathematical biophysics*, 14(4), 375-383.
- [37] **Erdos, P., & R&WI, A.** (1959). On random graphs I. *Publ. Math. Debrecen*, 6, 290-297.
- [38] **Erdos, P., & R&WI, A.** (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1), 17-60.
- [39] **Barabasi, A. L. L.** (2003). How everything is connected to everything else and what it means for business, science and everyday life. *New York: Plume*.
- [40] **Barabási, A. L., & Albert, R.** (1999). Emergence of scaling in random networks. *science*, 286(5439), 509-512.
- [41] **Girvan, M., & Newman, M. E.** (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), 7821-7826.

- [42] **Redner, S.** (2008). Networks: teasing out the missing links. *Nature*, 453(7191), 47.
- [43] **Allali, O., Magnein, C., & Latapy, M.** (2013). International link prediction: A new approach for predicting links in bipartite graphs. *International Data Analysis* 17, 5-25.
- [44] **Barabási, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T.** (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3-4), 590-614.
- [45] **Costa, L. D. F., Rodrigues, F. A., Travieso, G., & Villas Boas, P. R.** (2007). Characterization of complex networks: A survey of measurements. *Advances in physics*, 56(1), 167-242.
- [46] **Goldenberg, A., Zheng, A. X., Fienberg, S. E., & Airolidi, E. M.** (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2), 129-233.
- [47] **Chang, W. L., & Lin, T. H.** (2010, August). A cluster-based approach for automatic social network construction. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on* (pp. 601-606). IEEE.
- [48] **Newman, M. E., & Girvan, M.** (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2), 026113.
- [49] **Trudeau, R. J.** (2013). *Introduction to graph theory*. Courier Corporation.
- [50] **Gould, R.** (1988). *Graph Theory*. The Benjamim.
- [51] **Deo, N.** (2017). *Graph theory with applications to engineering and computer science*. Courier Dover Publications.
- [52] **Saxena, A., Malik, V., & Iyengar, S. R. S.** (2015). Estimating the Degree Centrality Ranking of a Node. *arXiv preprint arXiv:1511.05732*.
- [53] **Watts, D. J., & Strogatz, S. H.** (1998). Collective dynamics of ‘small-world’ networks. *nature*, 393(6684), 440.
- [54] **Newman, M. E., Watts, D. J., & Strogatz, S. H.** (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1), 2566-2572.
- [55] **Bütün, E., Kaya, M., & Alhajj, R.** (2017, July). A Supervised Learning Method for Prediction Citation Count of Scientists in Citation Networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 952-958). ACM.
- [56] **Ferreri, L., Venturino, E., & Giacobini, M.** (2011, April). Do diseases spreading on bipartite networks have some evolutionary advantage?. In *European*

Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (pp. 141-146). Springer, Berlin, Heidelberg.

- [57] **Zhang, Y., Tao, C., Jiang, G., Nair, A. A., Su, J., Chute, C. G., & Liu, H.** (2014). Network-based analysis reveals distinct association patterns in a semantic MEDLINE-based drug-disease-gene network. *Journal of biomedical semantics*, 5(1), 33.
- [58] **Kaya, B., & Poyraz, M.** (2016). Unsupervised link prediction in evolving abnormal medical parameter networks. *International Journal of Machine Learning and Cybernetics*, 7(1), 145-155.
- [59] **Kaya, B., & Poyraz, M.** (2014). Supervised link prediction in symptom networks with evolving case. *Measurement*, 56, 231-238.
- [60] **Ge, M., Li, A., & Wang, M.** (2016). A bipartite network-based method for prediction of long non-coding RNA–protein interactions. *Genomics, proteomics & bioinformatics*, 14(1), 62-71.
- [61] **Linden, G., Smith, B., & York, J.** (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, (1), 76-80.
- [62] **Banerjee, S., Jenamani, M., & Pratihari, D. K.** (2017, April). Properties of a projected network of a bipartite network. In *Communication and Signal Processing (ICCSP), 2017 International Conference on* (pp. 0143-0147). IEEE.
- [63] **Newman, M. E.** (2003). The structure and function of complex networks. *SIAM review*, 45(2), 167-256.
- [64] **Chen, D., Lü, L., Shang, M. S., Zhang, Y. C., & Zhou, T.** (2012). Identifying influential nodes in complex networks. *Physica a: Statistical mechanics and its applications*, 391(4), 1777-1787.
- [65] **Volkova, S., & Hsu, W. H.** (2009). Link prediction in social networks. *KDD Lab Course Project*.
- [66] **Aslan, S., & Kaya, M.** (2017). Link prediction methods in bipartite networks. In *Computer Science and Engineering (UBMK), 2017 International Conference on IEEE*. 1095-1099.
- [67] **ONAT, Y. D. D. F., & ALİKILIÇ, Ö. A.** (2008). Sosyal ağ sitelerinin reklam ve halkla ilişkiler ortamları olarak değerlendirilmesi. *Journal of Yaşar University*, 3(9), 1111-1143.
- [68] **Folino, F., & Pizzuti, C.** (2010). A comorbidity-based recommendation engine for disease prediction.

- [69] **Folino, F., & Pizzuti, C.** (2012, September). Link prediction approaches for disease networks. In *International Conference on Information Technology in Bio-and Medical Informatics* (pp. 99-108). Springer, Berlin, Heidelberg.
- [70] **Gül, S., Kaya, M., & Kaya, B.** (2016, August). Predicting links in weighted disease networks. In *Computer and Information Sciences (ICCOINS), 2016 3rd International Conference on* (pp. 77-81). IEEE.
- [71] **Liang, Y., Huang, L., & Wang, Z.** (2017, August). Link prediction in social network based on local information and attributes of nodes. In *Journal of Physics: Conference Series* (Vol. 887, No. 1, p. 012043). IOP Publishing.
- [72] **Lü, L., & Zhou, T.** (2011). Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6), 1150-1170.
- [73] **Sarukkai, R. R.** (2000). Link prediction and path analysis using Markov chains1. *Computer Networks*, 33(1-6), 377-386.
- [74] **Zhu, J., Hong, J., & Hughes, J. G.** (2002). Using markov chains for link prediction in adaptive web sites. In *Soft-Ware 2002: Computing in an Imperfect World* (pp. 60-73). Springer, Berlin, Heidelberg.
- [75] **Popescul, A., & Ungar, L. H.** (2003, August). Statistical relational learning for link prediction. In *IJCAI workshop on learning statistical models from relational data* (Vol. 2003).
- [76] **Guo, F., Yang, Z., & Zhou, T.** (2013). Predicting link directions via a recursive subgraph-based ranking. *Physica A: Statistical Mechanics and its Applications*, 392(16), 3402-3408.
- [77] **Zhou, T., Lü, L., & Zhang, Y. C.** (2009). Predicting missing links via local information. *The European Physical Journal B*, 71(4), 623-630.
- [78] **Grabowski, A., Kruszewska, N., & Kosiński, R. A.** (2008). Dynamic phenomena and human activity in an artificial society. *Physical Review E*, 78(6), 066110.
- [79] **Menon, A. K., & Elkan, C.** (2011, September). Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 437-452). Springer, Berlin, Heidelberg.
- [80] **Airolidi, E. M., Blei, D. M., Fienberg, S. E., Xing, E. P., & Jaakkola, T.** (2006, March). Mixed membership stochastic block models for relational data with application to protein-protein interactions. In *Proceedings of the international biometrics society annual meeting* (Vol. 15).
- [81] **Fire, M., Puzis, R., & Elovici, Y.** (2013). Link prediction in highly fractional data sets. In *Handbook of computational approaches to counterterrorism* (pp. 283-300). Springer, New York, NY.

- [82] **Benchettara, N., Kanawati, R., & Rouveirol, C.** (2010, August). Supervised machine learning applied to link prediction in bipartite social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on* (pp. 326-330). IEEE.
- [83] **Pujari, M.** (2015). *Link Prediction in Large-scale Complex Networks (Application to bibliographical Networks)* (Doctoral dissertation, Université Sorbonne Paris Cité).
- [84] **Sett, N., Singh, S. R., & Nandi, S.** (2016). Influence of edge weight on node proximity based link prediction methods: An empirical analysis. *Neurocomputing*, 172, 71-83.
- [85] **Navarro, G.** (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1), 31-88.
- [86] **Kunegis, J., & Lommatzsch, A.** (2009, June). Learning spectral graph transformations for link prediction. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 561-568). ACM.
- [87] **Martínez, V., Berzal, F., & Cubero, J. C.** (2017). A survey of link prediction in complex networks. *ACM Computing Surveys (CSUR)*, 49(4), 69.
- [88] **Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J.** (2012, February). Effects of user similarity in social media. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 703-712). ACM.
- [89] **Akcora, C. G., Carminati, B., & Ferrari, E.** (2013). User similarities on social networks. *Social Network Analysis and Mining*, 3(3), 475-495.
- [90] **Newman, M. E.** (2001). Clustering and preferential attachment in growing networks. *Physical review E*, 64(2), 025102.
- [91] **Adamic, L. A., & Adar, E.** (2003). Friends and neighbors on the web. *Social networks*, 25(3), 211-230.
- [92] **Leicht, E. A., Holme, P., & Newman, M. E.** (2006). Vertex similarity in networks. *Physical Review E*, 73(2), 026120.
- [93] **Zeng, S.** (2016). Link prediction based on local information considering preferential attachment. *Physica A: Statistical Mechanics and its Applications*, 443, 537-542.
- [94] **Barabási, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T.** (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3-4), 590-614.

- [95] Lü, L., Jin, C. H., & Zhou, T. (2009). Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, 80(4), 046122.
- [96] Corbellini, A., Mateos, C., Godoy, D., Zunino, A., & Schiaffino, S. (2016). An Evaluation of Distributed Processing Models for Random Walk-Based Link Prediction Algorithms Over Social Big Data. In *New Advances in Information Systems and Technologies* (pp. 919-928). Springer, Cham.
- [97] Doerfel, M. L., & Moore, P. J. (2016). Digitizing Strength of Weak Ties: Understanding Social Network Relationships through Online Discourse Analysis. *Annals of the International Communication Association*, 40(1), 127-148.
- [98] Valverde-Rebaza, J., & de Andrade Lopes, A. (2013). Exploiting behaviors of communities of twitter users for link prediction. *Social Network Analysis and Mining*, 3(4), 1063-1074.
- [99] Qiu, B., Ivanova, K., Yen, J., & Liu, P. (2010, August). Behavior evolution and event-driven growth dynamics in social networks. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on* (pp. 217-224). IEEE.
- [100] Pavlov, M., & Ichise, R. (2007). Finding experts by link prediction in co-authorship networks. *FEWS*, 290, 42-55.
- [101] Scripps, J., Tan, P. N., Chen, F., & Esfahanian, A. H. (2008, December). A matrix alignment approach for link prediction. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on* (pp. 1-4). IEEE.
- [102] Clauset, A., Moore, C., & Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191), 98.
- [103] Kashima, H., & Abe, N. (2006, December). A parameterized probabilistic model of network evolution for supervised link prediction. In *Data Mining, 2006. ICDM'06. Sixth International Conference on* (pp. 340-349). IEEE.
- [104] Bulut, H., & Milli, M. (2016). İşbirlikçi filtreleme için yeni tahminleme yöntemleri. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 22(2), 123-128.
- [105] Aslan, S., & Kaya, M. (2018). Topic recommendation for authors as a link prediction problem. *Future Generation Computer Systems*, 89, 249-264.
- [106] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web* (pp. 285-295). ACM.
- [107] Allali, O., Magnien, C., & Latapy, M. (2011, April). Link prediction in bipartite graphs using internal links and weighted projection. In *Computer*

Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on (pp. 936-941). IEEE.

- [108] **Shibata, N., Kajikawa, Y., & Sakata, I.** (2012). Link prediction in citation networks. *Journal of the American society for information science and technology*, 63(1), 78-85.
- [109] **Cui, Y., Zhang, L., Wang, Q., Chen, P., & Xie, C.** (2016). Heterogeneous network linkage-weight based link prediction in bipartite graph for personalized recommendation. *Procedia Computer Science*, 91, 953-958.
- [110] **Hoseini, E., Hashemi, S., & Hamzeh, A.** (2012, March). Link prediction in social network using co-clustering based approach. In *Advanced Information Networking and Applications Workshops (WAINA), 2012 26th International Conference on*(pp. 795-800). IEEE.
- [111] **Wu, T., Yu, S. H., Liao, W., & Chang, C. S.** (2014, October). Temporal bipartite projection and link prediction for online social networks. In *Big Data (Big Data), 2014 IEEE International Conference on* (pp. 52-59). IEEE.
- [112] **Davis, J., & Goadrich, M.** (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240). ACM.
- [113] **Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T.** (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1), 5-53.
- [114] **Zweig, K. A., & Kaufmann, M.** (2011). A systematic approach to the one-mode projection of bipartite graphs. *Social Network Analysis and Mining*, 1(3), 187-218.
- [115] **Neal, Z.** (2013). Identifying statistically significant edges in one-mode projections. *Social Network Analysis and Mining*, 3(4), 915-924.
- [116] **Neal, Z.** (2014). The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks*, 39, 84-97.

ÖZGEÇMİŞ

Serpil ASLAN

Munzur Üniversitesi
Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü

E-Posta: saslan@munzur.edu.tr

- 1990** Elazığ'da doğdu.
- 2004-2008** Elazığ Mehmet Akif Ersoy Yabancı Dil Ağırlıklı Lisesini tamamladı.
- 2008-2012** Fırat Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümünden mezun oldu.
- 2012-2014** Fırat Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Kuramsal Temeller A.B.D' da Yüksek Lisans eğitimini tamamladı.
- 2013-devam** Munzur Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümünde Araştırma Görevlisi olarak çalışmaktadır.
- 2014-** Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Kuramsal Temeller Anabilim dalında Doktora çalışmasına başladı.