



**KARMAŞIK AĞLARDA BAĞLANTI VE AĞIRLIKLARIN
BİRLİKTE TAHMİN EDİLMESİ: BİLİM İNSANLARININ
ATIF SAYISININ TAHMİNİ**

Yük. Müh. Ertan BÜTÜN

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

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KARMAŞIK AĞLARDA BAĞLANTI VE AĞIRLIKLARIN BİRLİKTE
TAHMİN EDİLMESİ: BİLİM İNSANLARININ ATIF SAYISININ TAHMİNİ

DOKTORA TEZİ

Yük. Müh. Ertan BÜTÜN

(131129205)

Tezin Enstitüye Verildiği Tarih : 24 Nisan 2018

Tezin Savunulduğu Tarih : 15 Mayıs 2018

Tez Danışmanı : Prof.Dr. Mehmet KAYA



Diğer Jüri Üyeleri : Prof.Dr. Ali KARCI



Doç.Dr. Muhammed Fatih TALU



Doç.Dr. Bilal ALATAŞ



Doç.Dr. Taner TUNCER



NİSAN-2018

ÖNSÖZ

Bu tez çalışmamda sahip olduğu bilgi birikimiyle bana danışmanlık yapan ve katkılarıyla tezin bu aşamaya gelmesini sağlayan değerli hocam Sayın Prof. Dr. Mehmet KAYA'ya teşekkürlerimi sunarım. Tez çalışması için gerekli donanımsal araçların temin edilmesinde maddi desteğinden dolayı Fırat Üniversitesi Bilimsel Araştırma Projeleri Birimine teşekkürü borç bilirim.

Yoğun çalışma zamanlarımda manevi desteğini esirgemeyip her zaman yanımda olan başta sevgili eşim ve tüm aileme teşekkür ederim.

Ertan BÜTÜN
ELAZIĞ-2018

İÇİNDEKİLER

	<u>Sayfa No</u>
ÖNSÖZ	II
İÇİNDEKİLER.....	III
ÖZET	V
SUMMARY	VI
ŞEKİLLER LİSTESİ	VII
TABLolar LİSTESİ	VIII
KISALTMALAR.....	IX
SEMBOLLER LİSTESİ	X
1. GİRİŞ	1
1.1. Tezin Amacı ve Katkıları	1
1.2. Tezin Yapısı	2
2. KARMAŞIK AĞLARDA BAĞLANTI TAHMİNİ	4
2.1. Bağlantı Tahmini Probleminin Tanımı	6
2.2. Bağlantı Tahmini İçin Önerilen Bazı Yaklaşımlar	7
2.2.1. Benzerlik Tabanlı Bağlantı Tahmini Yöntemleri.....	8
2.2.1.1. Komşuluk Tabanlı Ölçütler.....	8
2.2.1.2. Yol Tabanlı Ölçütler	10
2.2.1.3. Rastgele Yürüyüş Tabanlı Ölçütler	12
2.2.2. Öğrenme Tabanlı Bağlantı Tahmini Yöntemleri	13
2.2.2.1. Nitelik Tabanlı Sınıflandırma Yöntemleri	13
2.2.2.2. İstatistiksel Ağ Tabanlı Yöntemler	14
2.2.2.3. Matris Faktörizasyonu Tabanlı Yöntemler	16
2.3. Bağlantı Tahmini Probleminin Zorlukları.....	16
3. TEZ ÇALIŞMASINDA KULLANILAN AĞ VERİ KÜMELERİ VE DEĞERLENDİRME ÖLÇÜTLERİ	19
3.1. Kullanılan Ağ Verileri	19
3.2. Değerlendirme Ölçütleri	20
4. BİLİM İNSANLARININ ATIF SAYISINI TAHMİN ETMEK İÇİN YENİ BİR BAĞLANTI TAHMİNİ YAKLAŞIMI	22
4.1. Zamansal Bağlantı Tahmini Ölçütü	24
4.1.1. Zamansal Olaylar	25
4.1.1.1. Doğrudan Zamansal Olaylar	25

4.1.1.2.	Dolaylı Zamansal Olaylar	27
4.1.1.3.	Önerilen Zamansal Bağlantı Tahmini Ölçütünün Hesaplanması.....	27
4.2.	Eğiticili Bağlantı Tahmini.....	31
4.3.	Deneyler	32
4.3.1.	Deney Konfigürasyonları	32
4.3.2.	Test Veri Kümeleri ve Ağ Modeli	32
4.3.3.	Deney Sonuçları.....	33
4.3.3.1.	Deney 1: Bilim İnsanlarının Atıf Sayısı Tahmini	34
4.3.3.2.	Deney 2: Yeni Bağlantı Tahmini	37
4.4.	Bölüm Değerlendirmesi ve Sonuçlar	40
5.	KOMŞULUK TABANLI ÖLÇÜTLERİN YÖNLÜ, AĞIRLIKLI VE ZAMANSAL AĞLAR İÇİN GELİŞTİRİLMESİ.....	42
5.1.	Klasik Komşuluk Tabanlı Ölçütlerin Ağırlıklı ve Zamansal Olarak Genişletilmesi.....	44
5.2.	Komşuluk Tabanlı Ölçütlerin Yönlü Ağlar İçin Geliştirilmesi	47
5.3.	Eğiticili Bağlantı Tahmini.....	50
5.4.	Deneyler	50
5.4.1.	Deney Konfigürasyonları	50
5.4.2.	Test Veri Kümeleri ve Ağ Modeli	51
5.4.3.	Deney Sonuçları.....	52
5.4.3.1.	Deney 1: Komşuluk Tabanlı Ölçütlerin EXT_NLPM Ölçütleri ile Yönlü, Ağırlıklı ve Zamansal Ağlarda Karşılaştırılması	52
5.4.3.2.	Deney 2: EXT_NLPM ile TC'nin Yönlü Ağlarda Karşılaştırılması	64
5.5.	Bölüm Değerlendirmesi ve Sonuçlar	66
6.	SONUÇLAR VE ÖNERİLER	69
	KAYNAKLAR	71
	ÖZGEÇMİŞ.....	79

ÖZET

Birçok farklı alanda, özellikle de internet ortamındaki sosyal ağlarda, varlıklar arasındaki ilişkiler oldukça karmaşık ağlar oluşturur. Son zamanlarda karmaşık ağlarda ağ analizi ve veri madenciliği araştırmacıların ilgisini çekmiştir. Karmaşık ağlarda bağlantı tahmini de oldukça ilgi gören araştırma konularından biridir.

Bu çalışmada karmaşık ağlarda bağlantı tahmini için iki yöntem önerilmiştir. Birinci yöntemin amacı bilim insanlarının atıf sayısını tahmin etmektir. Önerilen yöntemde bilim insanlarının atıf sayısı tahmini problemi, atıf ağlarında bağlantı tahmini problemi olarak formüleleştirilmiştir. Atıf ağlarının zaman içindeki evrimi boyunca artan/azalan eğilimlerin dikkate alındığı zamansal bir bağlantı tahmini ölçütü önerilmiştir. Önerilen bağlantı tahmini ölçütü yönlü, ağırlıklı ve zamansal ağlarda bağlantıların ağırlıklarıyla beraber tahmin edildiği ilk çalışmadır. Atıf ağlarında yapılan test sonuçları önerilen yöntemin atıf sayısı tahmini için doğruluğunu gösterir. Önerilen bağlantı tahmini ölçütü ayrıca klasik bağlantı tahmini ölçütüyle de karşılaştırılmış ve önerilen ölçütün etkili bir bağlantı tahmini ölçütü test sonuçlarında görülmüştür.

İkinci yöntemin amacı yönlü ağlarda bağlantıların yön bilgisinin bağlantı oluşumundaki rolü dikkate alınarak komşuluk tabanlı bağlantı tahmini ölçütlerinin doğruluğunu artırmaktır. Literatürdeki bağlantı tahmini ölçütlerinin çoğunda bağlantıların yön bilgisinin bağlantı oluşumundaki rolü dikkate alınmamıştır. Bunun için bu çalışmada klasik komşuluk tabanlı bağlantı tahmini ölçütlerinin yönlü ağ motifleri kullanılarak hesaplandığı genel bir yöntem önerilmiştir. Yönlü ağlarda yapılan test sonuçları önerilen yöntemin komşuluk tabanlı bağlantı tahmini ölçütlerinin doğruluğunu dikkate değer bir şekilde artırdığını göstermiştir.

Anahtar Kelimeler: Bağlantı Tahmini, Sınıflandırma, Sosyal Ağlar, Atıf Ağları, Dinamik Ağlar, Yönlü ve Ağırlıklı Ağlar.

SUMMARY

Predicting of Links and Weights Together in Complex Networks: Prediction of Citation Count of Scientists

Interactions and collaborations between entities in many different areas, especially social networks in the Internet, create more complex networks. In recent times, network analysis and data mining in complex networks have attracted the attention of researchers. Link prediction in complex networks is also one of the most interesting research topics.

In this study, two methods were proposed for link prediction in complex networks. The aim of the first method is to predict citation count of scientists. In the proposed method, predicting citation count of scientists problem has been formulated as a link prediction problem in citation networks. A temporal link prediction metric has been proposed that takes into account upward/downward trends throughout the evolution of citation networks over time. The proposed link prediction approach is the first study that predicts links with its weights in the directed, weighted, and temporal networks. The experimental results on citation networks show the accuracy of the proposed method to predict citation count of scientists. The proposed link prediction metric was also compared with the classical link prediction metric, and it has been shown that the proposed measure is an effective link prediction metric in the test results.

The aim of the second method is to increase the accuracy of neighborhood-based link prediction metrics by considering the role of link direction information in link formation in the directed networks. The role of link direction information in link formation has not been considered in the majority of link prediction metrics in the literature. For this purpose, in this study, a general method is proposed in which the classical neighborhood based link prediction metrics are calculated by using on directional network motifs. Test results on directed networks have shown that the proposed method considerably improves the accuracy of neighborhood-based link prediction metrics.

Keywords: Link Prediction, Classification, Social Networks, Citation Networks, Dynamic Networks, Directed and Weighted Networks.

ŞEKİLLER LİSTESİ

	<u>Sayfa No</u>
Şekil 2.1. ACM, IEEE ve Elseiver’de bağlantı tahmini ile ilgili çalışmaların sayısı	5
Şekil 2.2. Bağlantı tahmini problemi için bir örnek.....	6
Şekil 2.3. Bağlantı tahmini yaklaşımlarının genel çerçevesi	7
Şekil 3.1. HEP-Th ağı şeması	19
Şekil 4.1. Örnek bir ağ	30
Şekil 4.2. Ağ modeli için örnek bir gösterim.....	33
Şekil 4.3. RF ile elde edilen sonuçların ROC eğrileri.....	39
Şekil 5.1. Açık üçlü ağ motifi	46
Şekil 5.2. Yönlü (a-c) ve yönsüz (d) bağlantılardan oluşan bir kaç OTP örneği	46
Şekil 5.3. OTP türleri	47
Şekil 5.4. EXT_NLPM hesaplamasının gösterimi için örnek bir ağ	49
Şekil 5.5. DBLP’de en iyi EXT_NLPM ölçütü sonucunun ve buna karşılık gelen komşuluk tabanlı ölçütü sonucunun ROC eğrileri.....	54
Şekil 5.6. DBLP’de komşuluk tabanlı ölçütlere göre en yüksek performans artışının sağlandığı sonuçların ROC eğrileri	55
Şekil 5.7. HEP-Th’de en iyi EXT_NLPM ölçütü sonucunun ve buna karşılık gelen komşuluk tabanlı ölçütü sonucunun ROC eğrileri.....	57
Şekil 5.8. HEP-Th’de komşuluk tabanlı ölçütlere göre en yüksek performans artışının sağlandığı sonuçların ROC eğrileri	58
Şekil 5.9. DBLP’de en iyi EXT_NLPM+ sonucunun ve buna karşılık gelen EXT_NLPM sonucunun ROC eğrileri	60
Şekil 5.10. DBLP’de EXT_NLPM’ye karşı EXT_NLPM+ ile en yüksek performans artışının sağlandığı sonuçların ROC eğrileri	61
Şekil 5.11. HEP-Th’de en iyi EXT_NLPM+ sonucunun ve buna karşılık gelen EXT_NLPM sonucunun ROC eğrileri	62
Şekil 5.12. HEP-Th’de EXT_NLPM’ye karşı EXT_NLPM+ ile en yüksek performans artışının sağlandığı sonuçların ROC eğrileri	63
Şekil 5.13. RF sınıflandırıcısında EXT_NLPM ve TC ölçüt sonuçlarının ROC eğrileri ..	67

TABLÖLAR LİSTESİ

	<u>Sayfa No</u>
Tablo 4.1. Doğrudan Zamansal Olay Denklemleri.....	26
Tablo 4.2. Dolaylı Zamansal Olaylar	27
Tablo 4.3. Örnek ağın skor değerleri	30
Tablo 4.4. Sınıf kategorileri.....	31
Tablo 4.5. Kullanılan nitelik kümeleri.....	31
Tablo 4.6. Örneklendirilmiş veri kümelerinin özellikleri.....	35
Tablo 4.7. Aminer AUROC sonuçları	36
Tablo 4.8. HEP-Th AUROC sonuçları	37
Tablo 4.9. Örneklendirilmiş veri kümelerinin özellikleri.....	38
Tablo 4.10. Aminer sonuçları	40
Tablo 4.11. HEP-Th sonuçları	40
Tablo 5.1. Komşuluk tabanlı ölçütlerin yönlü ağlarda ağırlıklı ve zamansal formları....	45
Tablo 5.2. Şekil 5.4'te $(vx, vy1)$ ve $(vx, vy2)$ düğüm çiftleri için klasik komşuluk tabanlı ve önerilen yönlü komşuluk tabanlı ölçüt değerleri	49
Tablo 5.3. HEP-Th ve DBLP atıf ağlarının özellikleri.....	51
Tablo 5.4. Örneklendirilmiş HEP-Th ve DBLP veri kümelerinin özellikleri	53
Tablo 5.5. DBLP ve HEP-Th veri kümelerinin AUROC sonuçları	53
Tablo 5.6. DBLP ve HEP-Th veri kümelerinin EXT_NLPM ve EXT_NLPM+ AUROC sonuçları	59
Tablo 5.7. Örneklendirilmiş veri kümelerinin özellikleri.....	65
Tablo 5.8. HEP-Th veri kümesinin AUROC sonuçları	66
Tablo 5.9. DBLP veri kümesinin AUROC sonuçları	66

KISALTMALAR

AA	: Adamic/Adar
AUROC	: Alıcı İşlem Karakteristikleri Altındaki Alan (Area Under Receiver Operating Characteristic)
CN	: Ortak Komşuluk (Common Neighbors)
CTP	: Kapalı Üçlü Örüntü (Closed Triad Pattern)
DBLP	: Bilgisayar Bilimi Bibliyografyası (Computer Science Bibliography)
FPR	: Yanlış Pozitif Oranı (False Positive Rate)
HEP-Th	: Yüksek Enerji Fiziği Teorisi (High Energy Physics Theory)
HD	: Hup Bastırılmış (Hup Depressed)
HP	: Hup Desteklenmiş (Hup Promoted)
IBk	: K-En Yakın Komşu
JC	: Jaccard Katsayısı (Jaccard Coefficient)
KNN	: Komşuların Sayısı (The Number of Neighbours)
LHN	: Leicht–Holme–Newman
LR	: Lojistik Regresyon (Logistic Regression)
MLP	: Çok Katmanlı Algılayıcılar (Multi Layer Perception)
OTP	: Açık Üçlü Örüntü (Open Triad Pattern)
RA	: Kaynak Paylaştırma (Resource Allocation)
RF	: Rastgele Orman (Random Forest)
ROC	: Alıcı İşlem Karakteristikleri (Receiver Operating Characteristic)
RT	: Rastgele Ağaç (Random Tree)
SA	: Salton
SO	: Sørensen
TC	: Üçlü Yakınlık (Triad Closeness)
TPR	: Doğru Pozitif Oranı (True Positive Rate)
WEKA	: Veri Madenciliği Yazılımı

SEMBOLLER LİSTESİ

V	: Ağdaki düğümlerin kümesi
E	: Ağdaki bağlantıların kümesi
$G_{t,t'}(V, E)$	: $[t, t']$ zaman aralığında yönlü, ağırlıklı ve zamansal ağ
n	: Zaman çerçevelerinin toplam sayısı
f	: Çerçeve kümesindeki f 'ninci çerçeve, $\{1, 2, \dots, f, \dots, n\}$
G_f	: f 'ninci çerçevedeki bağlantılardan oluşan bölünmüş ağ
$w(v_s, v_t, f)$	: G_f ağında v_s düğümünden v_t düğüme doğru olan bağlantının ağırlığı
$d(v_s, v_t, f)$	: G_f ağında v_s ile v_t arasındaki mesafe
$t(v_s, v_t)$	: v_s düğümünden v_t düğüme doğru olan bağlantının oluşma zamanı
$\Gamma(v_s)$	: v_s düğümünün komşularının kümesi
w	: zaman çerçevesinin genişliği
k_{v_x}	: v_x düğümünün komşularının eleman sayısı

1. GİRİŞ

Travers ve Milgram [1] tarafından 1967’te Amerika’da yapılan deneyde dünya küçüktür fenomeni ortaya atılmıştır. Yapılan bu deneyde Amerika’da bir grup insandan bir mesajı belirlenmiş bir kişiye ulaştırmaları istenmiştir. Deney sonucunda 4.4 ile 5.7 arasındaki ortalama kişi sayısı üzerinden bu mesajın bu insana ulaştığı görülmüştür. Backstrom vd. [2] çalışmasında 721 milyon aktif kullanıcı ve bunlar arasındaki 69 milyar arkadaşlık ilişkisinden oluşan tüm Facebook veri ağında yapılan hesaplamada kullanıcılar arasındaki ortalama mesafe 4.74 olarak ölçülmüştür. Yapılan bu deneyler varlıklar arasındaki ilişkilerin karmaşık ağlar oluşturduğunu gösterir. Sosyal ağlar (facebook, twitter, youtube, email ağları, vs.), web sayfaları arasındaki bağlantı ağları [3], bilimsel çalışmalar arasında oluşan biyografi ağları (atıf ağları [4], ortak yazarlık ağları [5], vs.), biyolojik ağlar (protein-protein etkileşimi ağları [6], hastalık-gen ağları [7], vs.) bu karmaşık ağlara birer örnektir. Karmaşık ağlardaki birçok saklı ve faydalı bilgiler biyologların, matematikçilerin, bilgisayar bilimcilerin, ekonomistlerin, sosyologların ve daha birçok bilim insanlarının ilgisini çekmiştir. İnternet kullanıcılarının ilgi alanlarının, alışverişlerinin, internet sayfalarındaki gezintilerinin analizi ile tavsiye edici sistemler [8-10], ağdaki aktörler arasında bağlantı tahmini [11-14], kullanıcıların davranışlarından örüntü çıkarımı [15-17], topluluk keşfi [18-20], terörist ağının analizi [21, 22] ile terörist gruplarına karşı önlem alma karmaşık ağlardaki çalışma konularından bazılarıdır. Bu tezde karmaşık ağlarda bağlantı tahmini problemi için iki yöntem önerilecektir.

1.1. Tezin Amacı ve Katkıları

Bu çalışmada karmaşık ağlarda bağlantı tahmini için iki yöntem önerilmiştir. Birinci yöntemin amacı bilim insanlarının atıf sayısının tahmin edilmesidir. Literatürde bilimsel etkilerin tahmini için yapılan çalışmaların çoğu bilimsel makale tabanlıdır. Bilim insanlarının bireysel olarak etkilerinin/performanslarının tahmini için az sayıda çalışma yapılmıştır. Bilim insanlarının bireysel etkilerini tahmin etmek, yapılacak bilimsel araştırmalar ve organizasyonlar için önemlidir. Önerilen yöntemde bilim insanlarının atıf sayısı tahmini problemi, bilimsel makaleler arasında oluşan karmaşık atıf ağlarında bağlantı tahmini problemi olarak ele alınmıştır. Atıf ağlarının zaman içindeki evrimi ile artan/azalan

eğilimlerin hesaplamaya katıldığı zamansal bir bağlantı tahmini ölçütü önerilmiştir. Önerilen bağlantı tahmini ölçütü yönlü, ağırlıklı ve zamansal ağlarda bağlantıların ağırlıklarıyla beraber tahmin edildiği ilk çalışmadır. Atıf ağlarında yapılan test sonuçları önerilen yöntemin atıf sayısı tahmini için doğruluğunu gösterir. Önerilen bağlantı tahmini ölçütü ayrıca klasik bağlantı tahmini ölçütüyle de karşılaştırılmış ve etkili bir bağlantı tahmini ölçütü olduğu test sonuçlarında gösterilmiştir.

Bu bölümde yapılan çalışmanın katkıları şöyle özetlenebilir:

- Bilim insanlarının atıf sayısı tahmini problemini ele alan az sayıdaki çalışmalara katkıda bulunulmuştur.
- Karmaşık ağlarda hem yerel hem de küresel topolojik yapılardan yararlanan, atıf ağlarının zaman içerisindeki evrimi boyunca artan/azalan eğilimlerini dikkate alan yeni bir zamansal bağlantı tahmini ölçütü önerilmiştir.
- Bilim insanlarının atıf sayısı tahmini problemi atıf ağlarında bir bağlantı tahmini problemi olarak formüle edilmiştir. Önerilen bağlantı tahmini yaklaşımında bağlantılar ağırlıklarıyla birlikte tahmin edilmiştir. Bu açıdan yapılan çalışma yönlü, ağırlıklı ve zamansal ağlarda bağlantıların ağırlıklarıyla beraber tahmin edildiği ilk çalışmadır.

Bağlantı tahmini için önerdiğimiz ikinci yöntemin amacı yönlü ağlarda bağlantıların yön bilgisinin bağlantı oluşumundaki rolü dikkate alınarak komşuluk tabanlı bağlantı tahmini ölçütlerinin doğruluğunun artırılmasıdır. Literatürdeki bağlantı tahmini ölçütlerinin çoğunda bağlantıların yön bilgisinin bağlantı oluşumundaki rolü dikkate alınmamıştır. Bunun için bu çalışmada klasik komşuluk tabanlı bağlantı tahmini ölçütlerin yönlü ağ motiflerine dayalı hesaplandığı genel bir yöntem önerilmiştir. Yönlü ağlarda yapılan test sonuçlarında önerilen yöntemin komşuluk tabanlı bağlantı tahmini ölçütlerinin doğruluğunu dikkate değer bir şekilde artırdığı görülmüştür.

1.2. Tezin Yapısı

Tezin sonraki bölümleri aşağıdaki gibi düzenlenmiştir:

- Bölüm 2’de öncelikle karmaşık ağlarda bağlantı tahmini probleminin tanımı, kapsamı ve uygulama alanları açıklanmıştır. Daha sonra literatürde yapılan çalışmalara kategorik olarak yer verilmiştir. Son olarak bağlantı tahmini probleminde yaşanan zorluklara değinilmiştir.

- Bölüm 3’te bu çalışmada kullanılan ağ veri kümelerinin özellikleri ve bağlantı tahmini için kullanılan değerlendirme ölçütleri verilmiştir.
- Bölüm 4’te bilim insanlarının atıf sayısı tahmini için bir bağlantı tahmini yöntemi önerilmiştir. Öncelikle yönlü, ağırlıklı ve zamansal ağlarda düğümlerin zaman içerisindeki eğilimlerini dikkate alan bir bağlantı tahmini ölçütü önerilmiştir. Önerilen bağlantı tahmini ölçütünde hem yerel hem de küresel topolojik niteliklerden faydalanılmıştır. Deney 1’de bağlantı tahmini ölçütü eğiticili bir bağlantı tahmini yönteminde nitelik olarak kullanılarak atıf sayısı tahmini yapılmıştır. Deney 2’de, önerilen bağlantı tahmini ölçütünün performansı geleneksel bir bağlantı tahmini ölçütü olan Adamic/Adar (AA) ölçütüyle geleneksel bağlantı tahmini testlerinde karşılaştırılmıştır.
- Bölüm 5’te, literatürdeki komşuluk tabanlı bağlantı tahmini ölçütlerin etkinliğinin artırılması amacıyla bu ölçütleri yönlü ağlar için genişleten genel bir yöntem önerilmiştir. Önerilen yöntem yönlü bağlantılardan oluşan ağ motiflerine dayalıdır. Ayrıca bağlantıların ağırlık ve zaman bilgileri de dikkate alınmıştır. Önerilen yöntem ile bağlantıların yön bilgisini de dikkate alan komşuluk tabanlı ölçütlerin performansı Deney 1’de test edilmiş ve dikkate değer bir şekilde performans artışının sağlandığı görülmüştür. Deney 2’de, önerilen yöntemle geliştirilen komşuluk tabanlı ölçütlerin performansı ağırlık ve zaman bilgilerini dikkate almayan yönlü ağlar için önerilmiş bir başka bağlantı tahmini ölçütü olan Üçlü Yakınlık Ölçütü (Triad Closeness, TC) ile test edilmiştir.
- Bölüm 6’da bu tezde elde edilen sonuçlar değerlendirilmiş ve gelecek araştırmalar için bazı önerilerde bulunulmuştur.

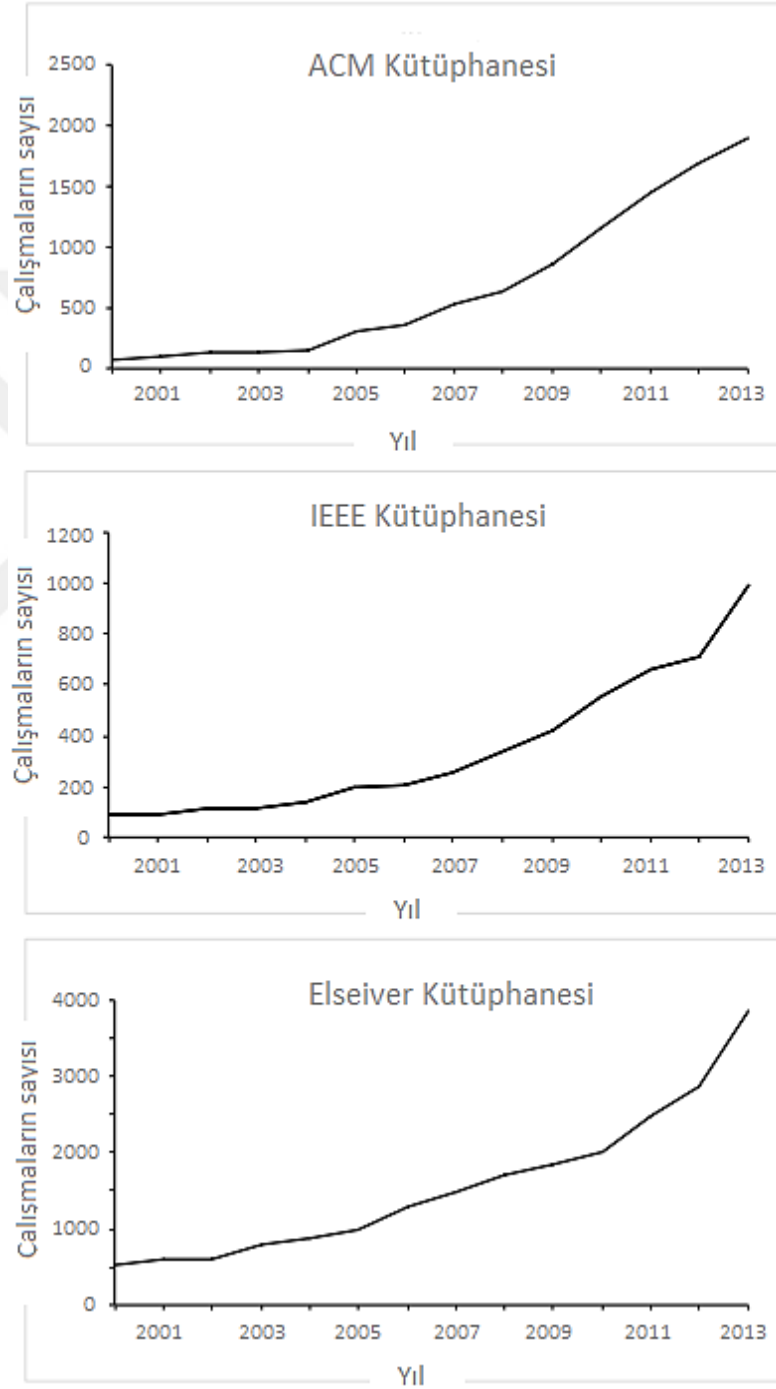
2. KARMAŞIK AĞLARDA BAĞLANTI TAHMİNİ

Bu bölümde karmaşık ağlarda bağlantı tahmini problemine ve önerilen bazı yaklaşımlara değinilecektir. İnternet teknolojilerinin hızla gelişmesi insanların çevrimiçi iletişim kurmalarını çevrimiçi iş birliktelikleri oluşturmalarını gittikçe artırmıştır. Bu sayede Facebook, Twitter gibi sosyal ağlar, e-ticaret ağları, bilim dünyasındaki çalışmalar arasında oluşan bibliyografik ağlar gibi çok büyük miktarda veriler ve bunlar arasındaki ilişkilerden oluşan oldukça karmaşık yapılar oluşmuştur. Birçok karmaşık ağ için ağ verisinin tümüne erişmek çoğu zaman mümkün değildir. Dolayısıyla erişilebilen karmaşık ağların çoğu; eksik, tamamlanmamış verilerden oluşur. Ayrıca karmaşık ağlar oldukça dinamik yapılardır. Bağlantıların ağa eklenmesi veya bağlantıların ağdan kaybolması sıklıkla olur. Karmaşık ağların eksik ve dinamik yapıda olmalarından dolayı eksik bağlantıların veya daha sonra eklenecek bağlantıların tahmini zor bir iştir. Karmaşık ağlarda bağlantı tahmininin birçok uygulama alanı vardır. Sosyal ağlarda yeni arkadaşlar tavsiye edilebilir [23], iş dünyasında ya da bilim dünyasında birlikte çalışılabilecek kişiler önerilebilir [24-26], teröristlerin etkileşimlerini gösteren bir ağda şüpheli kişiler arasında oluşabilecek ilişkiler önceden sezilebilir [27,28], web sayfaları arasında bağlantı tahmini yapılabilir [29], e-ticarette tüketicilere kişiselleştirilmiş ürün önerileri sunulabilir [30,31], dünya ticaret ağında ticari hareketler tahmin edilebilir [32], biyoinformatik alanında protein-protein [33] ve hastalık-gen etkileşimleri tahmin edilebilir [34].

Karmaşık ağların analizi ve karmaşık ağlarda bilgi keşfi son yıllarda fizikçiler, biyologlar, matematikçiler, bilgisayar bilimciler, ekonomistler, sosyologlar gibi birçok farklı disiplindeki araştırmacıların ilgisini çekmiştir. Wang vd. [35] tarafından bilgisayar bilimlerinde bilinen üç önemli kütüphane (ACM, IEEE, Elseiver) arşivinde 2000-2013 yılları arasında “bağlantı tahmini sosyal ağ” (“link prediction social network”) anahtar kelimeleri ile yapılan arama sonucunda çıkan çalışma sayıları Şekil 2.1’de verilmiştir. Şekil 2.1’de de görüldüğü gibi bağlantı tahmini önemli bilgisayar bilimleri kütüphanelerinde son yıllarda ilgi görmüş bir konudur.

Hasan ve Zaki tarafından [36] bağlantı tahmini probleminin, ağ gösteriminin olmadığı ilişkisel veri içeriklerinde ve internet alanında daha önceden çalışıldığı belirtilmiştir [37, 38, 39]. Bu çalışmalarda nesnelerin birbirleri ile karmaşık bir şekilde ilişkili olduğu herhangi bir ilişkisel veri kümesi ele alınabilmekte, veri kümesi içindeki bir nesne çifti arasında bağlantılar ve türleri tahmin edilebilmektedir. İlişkisel olasılık modelleri [40], grafiksel

modeller [41], stokastik ilişkisel modeller [42, 43] bu çalışmalarda kullanılan ana modellerdir. Bu yaklaşımların avantajları, varlıkların niteliklerinin modele kolay entegre edilebilmesi ve kapsayıcı olmalarıdır. Dezavantajları ise genellikle karmaşık olmaları ve çok sayıda parametreye ihtiyaç duymalarıdır [36].

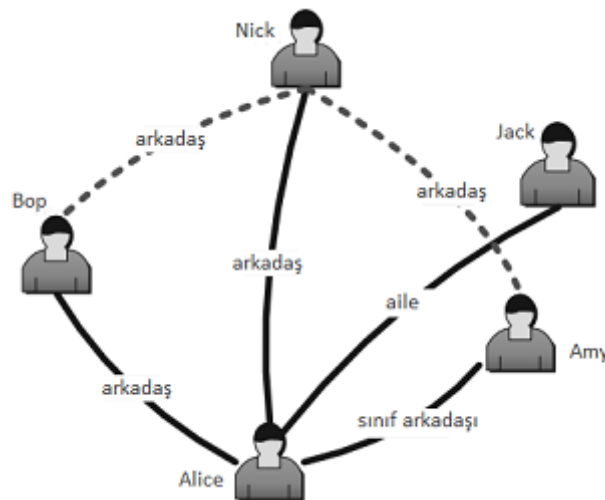


Şekil 2.1. ACM, IEEE ve Elsevier’de bağlantı tahmini ile ilgili çalışmaların sayısı [35].

Hasan ve Zaki çalışmasında [36] sosyal ağların gelişimi ya da evrimi konusundaki araştırmaların [44] bağlantı tahmini problemine çok yakın olduğu belirtilmiştir. Gelişim ya da evrim modelinde sosyal ağlarda dünya küçüktür fenomeni, güç kanunu dağılımı (power law distribution) gibi özellikler dikkate alınarak bağlantı tahmini yapılmaktadır. Farklılık olarak gelişim ya da evrim modelinde ağın genel özellikleri dikkate alınırken bağlantı tahmini probleminde aralarında bağlantı olup olmayacağı araştırılan düğümlerin durumları dikkate alınmaktadır [36].

2.1. Bağlantı Tahmini Probleminin Tanımı

Belirli bir $[t, t']$ zaman aralığında $G(V, E)$ bir karmaşık ağ olsun. Bağlantı tahminin amacı G ağında ileri bir $[t'', t''']$ zaman aralığında ($t'' > t'$) düğüm çiftleri arasında yeni oluşacak ya da kaybolacak bağlantıların veya içinde bulunulan $[t, t']$ zaman aralığında düğüm çiftleri arasında olan ancak bir şekilde gözden kaçan, görülemeyen bağlantıların tahmin edilmesidir [35]. Şekil 2.2’de bağlantı tahmini için örnek bir ağ verilmiştir. Bu ağda kişiler arasındaki düz bağlantılar $[t, t']$ zaman aralığındaki bağlantıları, kesik bağlantılar ise $[t'', t''']$ zaman aralığında oluşan bağlantıları göstermektedir. $[t, t']$ zaman aralığında Alice hem Bob ile hem de Nick ile arkadaştır. $[t'', t''']$ zaman aralığında Alice, Bob ile Nick’in arkadaş olmasına vesile olabilir. Bağlantı tahmininde amaç $[t'', t''']$ zaman aralığında oluşacak bağlantıların doğru tahmin edilmesidir.



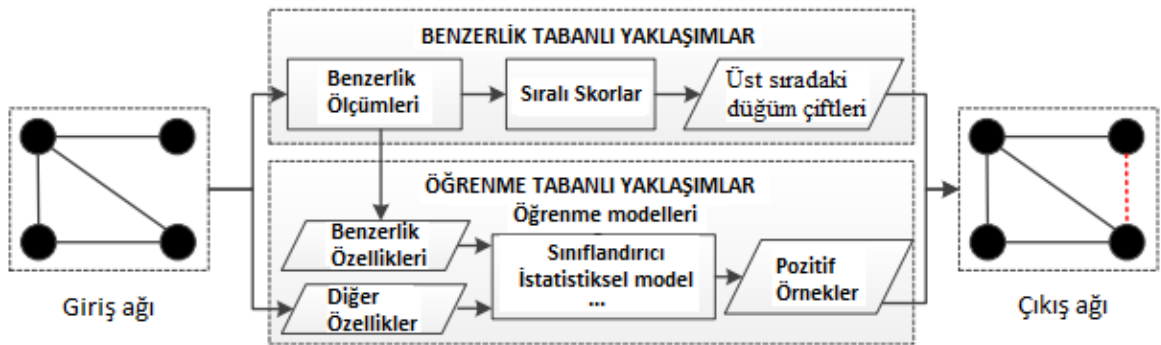
Şekil 2.2. Bağlantı tahmini problemi için bir örnek [35].

2.2. Bağlantı Tahmini İçin Önerilen Bazı Yaklaşımlar

Bağlantı tahmini yaklaşımları genel olarak iki kategoride ele alınabilir: benzerlik tabanlı yaklaşımlar, öğrenme tabanlı yaklaşımlar [35]. Bu iki yaklaşım literatürde sırayla eğitici-siz bağlantı tahmini ve eğitici bağlantı tahmini yöntemleri olarak ta geçmektedir.

Bağlantı tahmini yaklaşımlarının genel çerçevesi Şekil 2.3'te verilmiştir. Benzerlik tabanlı yaklaşımlarda giriş olarak verilen bir ağ verisinde aralarında bağlantı olma ihtimali bulunan tüm düğüm çiftlerinin skorları, önceden belirlenmiş bir benzerlik ölçütüne göre hesaplanır. Düğüm çiftlerinin skorları hesaplandıktan sonra bu skor değerleri azalan şekilde sıralanır. Daha sonra en üst sıradaki k tane düğüm çifti, tahmin edilen bağlantılar olarak çıkış ağında işaretlenir. k değerinin belirlenmesi ve sadece bir ölçütün dikkate alınması benzerlik tabanlı yaklaşımların dezavantajlarıdır.

Benzerlik tabanlı yöntemlerin dezavantajlarının üstesinden gelmek için öğrenme tabanlı bağlantı tahmini yöntemleri önerilmiştir. Öğrenme tabanlı yaklaşımlarda birden çok içerik veya ağ tabanlı benzerlik ölçütleri bir arada kullanılabilir. Belirlenen nitelikler giriş olarak verilen bir eğitim veri kümesinde hesaplanır. Belirlenen öğrenme modelinde aralarında bağlantı olan düğüm çiftleri ile bağlantı olmayan düğüm çiftlerinin nitelik değerlerini kullanarak bir tahmin edici model öğrenilir. Bu model aralarında bağlantı olacak düğüm çiftlerini pozitif, diğer düğüm çiftlerini negatif olarak etiketler. Öğrenme tabanlı yaklaşımlarda birden çok ağ tabanlı veya içerik tabanlı ölçütler kullanabildiğinden ve bunların hangisinin bağlantı oluşumunda daha belirleyici olduğu etkin bir şekilde öğrenebildiğinden bu yaklaşımlar genellikle benzerlik tabanlı yaklaşımlardan daha iyi sonuç verirler.



Şekil 2.3. Bağlantı tahmini yaklaşımlarının genel çerçevesi [35].

2.2.1. Benzerlik Tabanlı Bağlantı Tahmini Yöntemleri

Bu kategorideki yöntemlerde semantik veya topolojik tabanlı benzerlik ölçütleri kullanılır. Semantik ölçütlerde ağdaki düğümlerin içerik veya nitelik bilgilerinin belirlenen yöntemlerde kullanılmasıyla benzerlikler hesaplanır. Örneğin bir facebook arkadaşlık ağında insanların kişisel bilgileri, hobileri ile ilgili içerikler benzerlik için kullanılabilir [45]. Semantik tabanlı benzerlik ölçütler ağ yapılarıyla ilgili detaylı içerik bilgilerine ihtiyaç duydıklarından ve bu içeriklerin elde edilmesinin zorluğundan dolayı topolojik tabanlı ölçütler daha çok ilgi görmüştür.

Topolojik tabanlı ölçütlerde ağdaki düğümler ve düğümler arasındaki ilişkilerle oluşan yapısal özellikler kullanılır [46]. Ağdaki düğüm çiftleri arasındaki en kısa yolların sayısı topolojik tabanlı ölçütlere bir örnektir. Literatürde yaygın olarak kullanılan topolojik tabanlı ölçütler yaklaşım tarzlarına göre üçe ayrılır: komşuluk tabanlı, yol tabanlı ve rastgele yürüyüş tabanlı ölçütler. Sonraki kısımlarda bu kategorilerdeki ölçüt birimlerine değinilecektir.

2.2.1.1. Komşuluk Tabanlı Ölçütler

- **Adamic/Adar (Adamic/Adar, AA) İndeksi** [47]: Ağdaki düğümlerle az sayıda bağlantısı olan ve bu az sayıda bağlantısının aralarında ortak olduğu düğüm çiftlerinin bağlantı oluşturma ihtimali yüksektir.

$$AA(v_x, v_y) = \sum_{v_z \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{1}{\log(v_z)} \quad (2.1)$$

- **Ortak Komşuluk (Common Neighbors, CN) İndeksi** [48]: CN ölçütü, hesaplanmasının kolay olmasından dolayı en basit bağlantı tahmini ölçütlerinden biridir. v_x ve v_y düğümlerinin ortak komşularının sayısı ne kadar fazla ise aralarında bağlantı olma ihtimali o kadar yüksektir.

$$CN(v_x, v_y) = |\Gamma(v_x) \cap \Gamma(v_y)| \quad (2.2)$$

- **Hup Desteklenmiş (Hup Promoted, HP) İndeksi** [49]: Hup düğümlerine komşu olan düğüm çiftleri daha yüksek puanlandırılır. Hub düğümleri, düşük dereceye sahip düğümleri yüksek dereceye sahip düğümlere yönlendiren bir rol oynar.

$$HP(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{\min(k_{v_x}, k_{v_y})} \quad (2.3)$$

- **Hup Bastırılmış (Hup Depressed, HD) İndeksi** [50]: Bu ölçüt HP ölçüsüne benzer, ancak yüksek dereceye sahip düğümlerden etkilenir. Yüksek dereceye sahip herhangi bir düğüm bu ölçütte düşük puanlandırılır.

$$HD(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{\max(k_{v_x}, k_{v_y})} \quad (2.4)$$

- **Jaccard Katsayısı (Jaccard Coefficient, JC)** [51]: Bu ölçütte toplam komşu sayısına oranla daha çok sayıda ortak komşulara sahip düğüm çiftleri arasında bağlantı oluşma ihtimali yüksektir.

$$JC(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{|\Gamma(v_x) \cup \Gamma(v_y)|} \quad (2.5)$$

- **Leicht–Holme–Newman (Leicht–Holme–Newman, LHN) İndeksi** [52]: Bu ölçütte düğümlerin komşu sayılarının çarpımına oranla aralarında daha çok ortak komşu olan düğüm çiftleri yüksek puan elde ederler.

$$LHN(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{k_{v_x} \times k_{v_y}} \quad (2.6)$$

- **Kaynak Paylaştırma (Resource Allocation, RA) İndeksi** [50]: RA ölçütü AA ölçüsüne benzemektedir. Ancak RA ölçütünde ortak komşuları yüksek dereceye sahip olan düğüm çiftleri için daha düşük puan üretilir.

$$RA(v_x, v_y) = \sum_{v_z \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{1}{\log(v_z)} \quad (2.7)$$

- **Salton (SA) İndeksi** [53]: Salton ve McGill tarafından önerilen bu ölçüt, benzerlik ölçümlerinde yaygın olarak kullanılan kosinüs benzerliğine dayanmaktadır.

$$SA(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{\sqrt{k_{vx} \times k_{vy}}} \quad (2.8)$$

- **Sørensen (SO) İndeksi** [54]: SO ölçütü, bitki sosyolojisinde türlerin benzerliğine göre eşit gruplar oluşturmak için önerilmiştir. Karmaşık ağlarda düğümlerin benzerliklerini hesaplamak için de kullanılır. Düğümlerin derecelerinin toplamına oranla düğüm çiftlerinin ortak komşularının sayısıyla belirlenir.

$$SO(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{k_{vx} + k_{vy}} \quad (2.9)$$

2.2.1.2. Yol Tabanlı Ölçütler

Yol tabanlı ölçütler ağda iki düğüm arasında bağlantılar ile oluşan yollar üzerinden hesaplanırlar.

- **Yerel Yol (Local Path, LP):** Bu ölçütte [55] düğüm çiftleri arasındaki 2 ve 3 uzunluğundaki yollar bağlantı tahmini için kullanılır. Önerilen ölçüte göre sadece düğüm çiftlerinin komşularından ziyade ağda 3 birim uzaklıkta bulunan topolojik yapılar da bağlantı oluşumunda belirleyicidir. 3 birim uzaklıkta bulunan yapıların skora hangi ölçüde katkıda bulunacağı α parametresi ile belirlenir. LP ölçütü denklem 2.10'da verilmiştir. A^2 ve A^3 , düğümler arasında 2 ve 3 uzunluğunda mesafelerin gösterildiği komşuluk matrisleridir. Dolayısıyla LP ölçütü 2 ve 3 uzunluğundaki düğüm çiftlerini barındıran bir komşuluk matrisidir.

$$LP = A^2 + \alpha A^3 \quad (2.10)$$

- **Katz:** Katz [56] ölçütüyle ağdaki düğüm çiftleri arasındaki en kısa yolların sayısının toplamı hesaplanır. Ölçütün denklemi 2.11’de verilmiştir.

$$\text{Katz}(v_x, v_y) = \sum_{l=1}^{\infty} \beta^l \cdot |yol_{v_x, v_y}^{(l)}| = \beta A + \beta A^2 + \beta A^3 + \dots \quad (2.11)$$

β sıfırdan büyük sabit değerden oluşan bir parametredir, l kaç uzunluğundaki yolların sayılacağını gösterir. Bu iki parametrenin β^l formunda kullanılmasıyla düğümler arasındaki uzun yolların skor değerine katkısı hafifletilir. $|yol_{v_x, v_y}^{(l)}|$, v_x ve v_y düğümleri arasında l uzunluğundaki yolların sayısının toplamıdır. Düğüm çiftleri arasında ağdaki tüm yolların sayısı hesaplandığından hesaplama zamanı yüksektir. Liben-Nowell ve Kleinberg [11] tarafından Katz ölçütünün bağlantıların ağırlıklı olduğu ağlar için de hesaplaması gösterilmiştir. Ağırlıklı ağlarda birbirine doğrudan bağlı olan düğüm çifti arasındaki yol uzaklığı için bir yerine ağırlık değeri kullanılmıştır.

- **İlişki Gücü Benzerliği (Relation Strength Similarity, RSS) [57]:** Ağırlıklı ağlar için de uygulanabilen asimetrik bir ölçüttür. Ölçütün hesaplanması v_x ve v_y düğümlerinin komşularının benzerlik derecesini gösteren bir ilişki gücü $R(v_x, v_y)$ fonksiyonuna dayanır. v_x düğümünden v_y düğümüne uzunluğu r ’den düşük L tane yol $p_1, p_2, \dots, p_L$ olsun. Bu yollardan herhangi bir p_l yolu K tane düğümden $z_1, z_2, \dots, z_{k-1}, z_k$ oluşsun. v_x düğümünden v_y düğümüne RSS ölçütünün hesaplanması denklem 2.12’de verilmiştir.

$$\text{RSS}(v_x, v_y) = \sum_{l=1}^L R_{p_l}^*(v_x, v_y) \quad (2.12)$$

$$R_{p_l}^*(v_x, v_y) = \begin{cases} \prod_{k=1}^K R(z_{k-1}, z_k) & K \leq r \\ 0 & \text{değilse} \end{cases} \quad (2.13)$$

2.2.1.3. Rastgele Yürüyüş Tabanlı Ölçütler

- **Ulaşma zamanı (Hitting time, HT)** [58]: v_x düğümünün başlangıç noktası olduğu rastgele bir yürüyüş başlatılır. Bu rastgele yürüyüş düğümlerin komşuları üzerinden adım adım devam ettirilir. v_y düğümüne erişilmesi için atılması gereken adım sayısı bu ölçütün skor değeridir. v_x 'ten v_y 'ye doğru başlatılan yürüyüşün az sayıda adımla tamamlanması bu düğüm çifti arasında bağlantı oluşma ihtimalinin yüksek olduğu anlamına gelir. Ölçüt simetrik olmadığından hesaplama hem v_x 'ten v_y 'ye hem de v_y 'den v_x 'e doğru rastgele yürüyüşler ile yapılır. Ulaşma zamanının denklemi 2.14'te verilmiştir.

$$HT(v_x, v_y) := -\left(H_{(v_x, v_y)} \cdot \pi_{v_y} + H_{(v_y, v_x)} \cdot \pi_{v_x}\right) \quad (2.14)$$

π_{v_y} ve π_{v_x} normalleştirme amacıyla kullanılan parametrelerdir.

- **Rooted PageRank** [11]: Ulaşma zamanı ölçütünde ağda çok uzun yollar şeklinde oluşan durumların da hesaplama katılması bir dezavantajdır. Rooted PageRank ölçütünde v_x ve v_y arasındaki rastgele yürüyüş, belirlenmiş bir olasılık parametresi ile yeniden başlatılır. Böylece çok uzun yolların tamamlanmasından kaçınılır. Bu yaklaşım web arama motorlarında arama sonuçlarının derecelendirilmesi için kullanılan PageRank [59] ölçütünden ilham alınarak önerilmiştir. Rastgele yürüyüş $[0,1]$ aralığında belirlenen α olasılığı ile başa döner, $1 - \alpha$ olasılığı ile o an bulunan düğümün komşu düğümlerinden birine rastgele geçer. A komşuluk matrisi kullanılarak düğümlerin komşularının sayısının gösterildiği D diyagonal matrisi oluşturulur. D diyagonal matrisinde v_x düğümünün komşularının sayısı $D[v_x, v_x] = \sum_{v_i} A[v_x, v_i]$ 'dir. Ölçütün formülü denklem 2.15'te verilmiştir.

$$RPR = (1 - \alpha)(I - \alpha D^{-1}A)^{-1} \quad (2.15)$$

- **SimRank** [60]: Bu ölçüte göre iki düğüm benzer komşulara sahip olduğu ölçüde benzerdir. Komşuların benzerliği özyinelemeli olarak hesaplanır. SimRank ölçütü denklem 2.16'ya göre hesaplanır.

$$\text{SimRank}(v_x, v_y) = \gamma \cdot \frac{\sum_{a \in \Gamma(v_x)} \sum_{b \in \Gamma(v_y)} \text{simRank}(v_x, v_y)}{|\Gamma(v_x)| \cdot |\Gamma(v_y)|} \quad (2.16)$$

2.2.2. Öğrenme Tabanlı Bağlantı Tahmini Yöntemleri

2.2.2.1. Nitelik Tabanlı Sınıflandırma Yöntemleri

Bu kategorideki yaklaşımlarda bağlantı tahmini problemi ikili sınıflandırma problemi olarak ele alınır. Test veri kümesinde aralarında bağlantı bulunan düğüm çiftleri pozitif sınıfa, bağlantı bulunmayan düğüm çiftleri ise negatif sınıfa girerler. Belirlenen nitelikler düğüm çiftleri için eğitim veri kümelerinde hesaplanır. Hesaplanan nitelik değerleri Destek Vektör Makineleri, Karar Ağaçları gibi sınıflandırıcı modellerinde kullanılarak düğümler arasında bağlantı olup olmayacağı tahmini yapılır. Etkili bir bağlantı tahmini için karmaşık ağlardaki düğümler için uygun niteliklerin belirlenmesi bu yaklaşımların en kritik noktasıdır. Düğüm tabanlı, topolojik tabanlı ve sosyal teori tabanlı nitelikler yaygın olarak kullanılır.

Scellato vd. [61] konum tabanlı sosyal ağlarda sosyal tabanlı, konum tabanlı ve küresel tabanlı nitelikleri dikkate alan eğitici bir bağlantı tahmini yöntemi önermişlerdir. Konum tabanlı nitelik olarak iki kullanıcının aynı mekânda kaç kez bulunduğu bilgisi ve bu bilginin kosinüs benzerliği bilgisi; sosyal tabanlı nitelik olarak iki kullanıcının ortak arkadaşları üzerinden hesaplanan ve Bölüm 2’de komşuluk tabanlı ölçütler olarak geçen CN, JC ve AA ölçütleri; küresel tabanlı nitelik için iki kullanıcının sosyal ağda aralarındaki bağlantı mesafesi vb. bilgiler kullanılmıştır.

Ichise vd. [62] tarafından bilim dünyasındaki ortak yazarlık ağlarında topolojik niteliklerin yanı sıra bilimsel makalelerin özet, başlık gibi topolojik olmayan içerik tabanlı bilgilerin de nitelik olarak kullanılarak daha etkili bağlantı tahmini yapılabileceği gösterilmiştir. Öğrenme tabanlı yaklaşımlarda birçok nitelik bilgisi olduğundan etkili bir bağlantı tahmini yapmak için hangi niteliklerin belirleneceği en önemli noktadır.

Scripps vd. [63] hangi nitelik bilgisinin bağlantı tahmini için daha belirleyici olduğunu otomatik olarak belirleyen matris hizalamasına (matrix alignment) dayanan bir bağlantı tahmini yöntemi önermiştir. Önerilen yöntemde ağın komşuluk matrisi düğüm tabanlı ve topolojik tabanlı niteliklerden hesaplanan bir benzerlik matrisiyle hizalanmıştır.

De Sá ve Prudêncio [64] bağlantıların ağırlık bilgilerinin öğrenme tabanlı bağlantı tahmini yaklaşımlarla bağlantı tahmininin doğruluğunu artırdığını ortaya koyan bir çalışma

yapmışlardır. Bağlantı tahmini için kullanılan klasik ölçütlerin birçoğu ağırlıklı ağlar için yeniden düzenlenmiştir.

Lichtenwalter vd. [65] tarafından bağlantı tahmini problemi için ağırlıklı/ağırlıksız, yönlü/yönsüz ağlarda çalışabilen öğrenme tabanlı genel bir yaklaşım önerilmiştir. Çalışmada ayrıca yöntemin genelliği, varyans azaltma, topolojik problemler, derece dengesizliği ve örnekleme yaklaşımı konuları da incelenmiştir. Çalışmada PropFlow olarak isimlendirilen bir bağlantı tahmini ölçütü önerilmiştir. Bu ölçüt, v_i düğümünden başlayıp v_j düğümünde biten l uzunluğundaki rastgele yürüyüşün olasılığına dayanmaktadır. Rastgele yürüyüşteki düğüm geçişlerinin olasılıkları düğümler arasındaki bağlantıların ağırlıkları ile doğru orantılıdır. Rooted Pagerank ölçütüne benzemektedir. Ancak l uzunluğu sınırı ile PropFlow ölçütünün hesaplanması Rooted Pagerank ölçütüne göre daha yerel ölçekte kalmakta ve böylece ağırlık kaynak düğümden çok uzak kısımlarındaki gürültüler azaltılmaktadır.

Gerçek ağlarda düğümlerin ve bağlantıların farklı türlerde olduğu heterojen yapılar oldukça görülmektedir. Bağlantı tahmini için yapılan çalışmaların çoğunda heterojen ağlarda farklı türdeki ilişkilerin birlikte bağlantı oluşumundaki rolü pek incelenmemiştir. Davis vd. [34] tarafından yapılan çalışmada heterojen ağlar için bağlantı tahmini yöntemi önerilmiştir. Önerilen yöntemde AA ölçütü heterojen ağlar için bağlantıların ağırlıkları dikkate alınarak genişletilmiştir. Bahsedilen çalışmaların dışında nitelik tabanlı sınıflandırma yöntemlerinin kullanıldığı birçok bağlantı tahmini çalışması [66-75] yapılmıştır.

2.2.2.2. İstatistiksel Ağ Tabanlı Yöntemler

İstatistiksel ağ tabanlı yöntemlerde düğümleri birbirine bağlayan bağlantılarla olasılık değerleri atanır. Atanacak olasılık değerleri çoğunlukla topolojik benzerliklere veya düğümler arasında rastgele yürüyüşteki geçiş olasılıklarına göre hesaplanır. Bu olasılık değerleri hesaplanarak olasılık ağları oluşturulur. Daha sonra bu ağlar öğrenme tabanlı bağlantı tahmini yaklaşımlarında kullanılır.

Wang vd. [76] tarafından bağlantı tahmini probleminin çözümü için düğümlerin birlikte görülme olasılığını, topolojik nitelikleri ve anlamsal nitelikleri kullanan bir bağlantı tahmini yöntemi önerilmiştir. Birlikte görülme olasılığının (iki düğüm arasındaki bağlantı olasılığı) hesaplanması için Markov Rastgele Alanları (Markov Random Fields, MRF) kullanılarak yerel bir olasılıksal ağ modeli önerilmiştir. v_x ve v_y düğümleri arasında bağlantı oluşup oluşmayacağı tahmini için öncelikle düğüm çiftinin ortak komşularının kümesi bulunur.

Ardından bu kümedeki elemanlar boyunca sahip olunan nitelikler kümesi seçilir. Seçilen bu nitelik kümesi önerilen yerel olasılıksal modelin eğitilmesi için eğitim verilerinde kullanılır. Böylece düğüm çiftinin birlikte görülme olasılığı tahmin edilir. Elde edilen bu tahmin değeri ve diğer belirlenen nitelikler Lojistik Regresyon sınıflandırıcısında uygulanarak bağlantı tahmini yapılır.

Soares ve Prudencio [77] geleneksel bağlantı tahmini yaklaşımlarının ağıın zaman içindeki evrimini dikkate almada yetersiz olduğunu, bu yaklaşımlarda ağıın şu anki durumunun statik olarak ele alındığını, hangi bağlantının ne zaman oluştuğu ya da ne zamandır devam ettiği bilgisine bakılmadığını belirtmişlerdir. Çalışmada [77] bu kısıtların üstesinden gelmek için ağıın zaman içinde nasıl değiştiği araştırılmıştır. Bağlantı tahmini için önerilen algoritmada öncelikle veri ağları belirli zaman periyotlarına göre alt ağlara bölünmüştür. Daha sonra aralarında bağlantı olmayan tüm düğüm çiftlerinin benzerlikleri belirlenen ölçütlerle bölünmüş ağlarda hesaplanarak zaman serileri oluşturulmuştur. Oluşturulan zaman serileri, belirlenen tahmin modellerinde (forecasting model) kullanılarak düğüm çiftlerinin sonraki zaman dilimindeki benzerlikleri tahmin edilmiştir. Son olarak eğitici ve eğitici olmayan metotlarla bağlantı tahmini yapılmıştır. Daha etkin bağlantı tahmini için zaman serilerinin kullanıldığı başka çalışmalar da [78-81] olmuştur.

Backstrom ve Leskovec [82] eğitici rastgele yürüyüş tabanlı bir bağlantı tahmini yöntemi önermişlerdir. Geleneksel PageRank tüm bağlantıların geçiş olasılıklarının aynı olduğu varsayımına göre hesaplanırken bu çalışmada önerilen eğitici rastgele yürüyüş modeli ile her bağlantı için farklı geçiş olasılıkları belirleyecek bir fonksiyon öğrenilir. Öğrenilen bu fonksiyon sayesinde rastgele yürüyüşün ağıdaki diğer düğümlere göre hedef düğümleri ziyaret etme olasılığı daha yüksek olur. Bu fonksiyon düğümlerin ve düğümler arasındaki bağlantıların nitelikleri kullanılarak hesaplanır. Böylece bağlantı tahmini için ağıdaki düğüm ve bağlantıların yapısal niteliklerinden birlikte faydalanılır. v_x düğümünden diğer düğümlere doğru oluşacak yeni bağlantıların tahmini şöyle yapılır. Bağlantıların geçiş olasılıkları önerilen fonksiyonla hesaplanır. Ardından v_x düğümünden bir rastgele yürüyüş başlatılır ve üzerinden geçilen düğümlere bir olasılık değeri atanır. En yüksek olasılık değerine sahip düğümlerle v_x düğümü arasında bağlantı oluşacağı tahmini yapılır.

2.2.2.3. Matris Faktörizasyonu Tabanlı Yöntemler

Menon vd. [83] bağlantı tahmini probleminin çözümü için eğitici bir matris faktörizasyon yaklaşımı önermişlerdir. Önerilen model ile eğitici yaklaşımlarından daha etkili bağlantı tahmini yapmak için ağın topolojik yapılarındaki gizli nitelikler (latent features) öğrenilmiştir. Bu gizli niteliklerin düğüm ve bağlantıların belli olan nitelikleriyle nasıl entegre edileceği gösterilmiştir. Son olarak bağlantı tahminindeki sınıf dengesizliği problemi için yeni bir yaklaşım sunulmuştur. Önerilen modelin stokastik gradyan iniş (stochastic gradient descent) yaklaşımı ile optimizasyonu yapılarak büyük ağlarda ölçeklendirilebilirliği sağlanmıştır.

Dunlavy vd. [14] tarafından zamansal ikili (bipartite) ağlarda bağlantı tahmini için matris ve tensor faktörizasyonu tabanlı bir yöntem önerilmiştir. Farklı yıllardaki verilerin tek bir matris halinde gösterimi için ağırlık tabanlı bir yöntem sunulmuştur. Yol tabanlı ölçütlerden Katz ölçütü ikili ağlar için genişletilmiştir. Ağda üç boyutlu yapıdaki zamansal bağlantı verilerinin tensör ayrıştırma yöntemleriyle bağlantı tahmini için nasıl kullanışlı hale getirildiği gösterilmiştir. Daha sonra bu verilerde sezgisel tabanlı ve tahmin edici tabanlı yöntemlerle bağlantı tahmini yapılmıştır.

Gao vd. [84] tarafından matris faktörizasyonu ve grafik düzenleyici (graph regularization) tekniklere dayalı yeni bir birleşik model önerilmiştir. Önerilen model ile içerik ve yapısal bilgiler entegre edilerek zaman evrimi boyunca ağda oluşan bağlantı örüntüleri belirlenmiştir. Önerilen yöntem düğümlerdeki gizli nitelikleri etkili bir şekilde öğrenebilmekte ve büyük ağlara uygulanabilmektedir.

2.3. Bağlantı Tahmini Probleminin Zorlukları

Bağlantı tahmini problemi için birçok çalışma yapılmasına karşın problemin bazı zorlukları hala çalışma gerektirmektedir. Bu zorluklar şöyle özetlenmiştir [35]:

- **Kaybolan bağlantıların tahmini:** Çoğu bağlantı tahmini çalışmasında ağda oluşacak yeni bağlantılar tahmin edilmeye çalışılmış, ağda kaybolan bağlantıların tahmini için ise az sayıda çalışma yapılmıştır. Ağda kaybolacak bağlantıların tahmin işi, ağa eklenecek bağlantıların tahmin işinin tam olarak tersi değildir. Çünkü bağlantıların oluşum mekanizmalarıyla bağlantıların kaybolma mekanizmaları birbirinden

farklıdır. Bu yüzden kaybolan bağlantıların tahmini problemi, üzerinde çalışma gerektiren bir problemdir.

- **Düğümlerin dinamik olduğu ağlarda bağlantı tahmini:** Çoğu bağlantı tahmini yönteminde ağdaki düğümler statik olarak ele alınır ve durumlarının değişmeyecekleri varsayılır. Sabit olarak düşünülen düğümler arasında bağlantılar tahmin edilmeye çalışılır. Ancak bazı dinamik ağlarda bu durum geçerli değildir, düğümlerin ağa katılması veya ağdan ayrılması sıklıkla görülür. Örneğin Twitter ağında bazı kullanıcılar uzun bir süre aktif değildir. Böyle bir durumda bu kullanıcıların bağlantı tahmininde dikkate alınmaması gerekir. Bağlantı tahmini için önerilen çoğu yöntem düğümlerin dinamik olduğu ağlarda iyi sonuçlar vermeyebilir.
- **Sınıf dengesizliği problemi:** Gerçek karmaşık ağların büyük çoğunluğu çok seyrek, aralarında bağlantı olan düğüm çiftlerinin sayısı aralarında bağlantı olmayan düğüm çiftlerinin sayısına oranla çok azdır. Bu durum, sınıf dengesizliği problemine yol açar ve önerilen bir bağlantı tahmini yönteminin gerçek performansının doğru bir şekilde değerlendirilmesini zorlaştırır.
- **Sosyal teorilerden yararlanma:** Çoğu bağlantı tahmini yöntemi sosyal ağların yapısındaki topolojik nitelikleri dikkate alır. Bu durumun hem olumlu hem de olumsuz tarafı vardır. Olumlu tarafı önerilen yöntemin çoğu ağa uygulanabilir olmasıdır. Olumsuz tarafı ise sadece ağ yapısından yararlanması ve bağlantı tahmini performansının artırılması için sosyal teorilerden faydalanmamasıdır. Sosyal teoriler sosyal aktivitelerin, mekanizmaların anlaşılmasında kullanılabilir; böylece daha etkili bağlantı tahmini yaklaşımları sunulabilir.
- **Heterojen sosyal ağlarda bağlantı tahmini:** Geleneksel bağlantı tahmini yöntemlerinde ağlar homojen olarak ele alınır. Bu yöntemlerde düğümler ve bağlantılar tek tiptedir. Ancak birçok sosyal ağ farklı bağlantı ve düğüm türlerinden oluşmaktadır. Örneğin bilim dünyasında yapılan çalışmaların verilerinin bulunduğu bir biyografik ağda yazarlar, makaleler, dergiler birer düğüm; makaleler arasındaki atıflar, makalelerin yazar bilgileri, yayımlandıkları dergi bilgileri, ortak yazarlıklar birer bağlantı olabilir. Böyle bir ağda farklı düğüm ve bağlantı türleri arasında bağlantı tahmini gerekebilir veya bu farklı türlerin bağlantı oluşum mekanizmasına etkilerinden yararlanılabilir. Heterojen ağlarda bağlantı tahmini yapan çalışmalar [34, 85-89] olsa da daha etkin bağlantı tahmini için üzerinde araştırma yapmaya değer bir potansiyel barındırmaktadırlar.

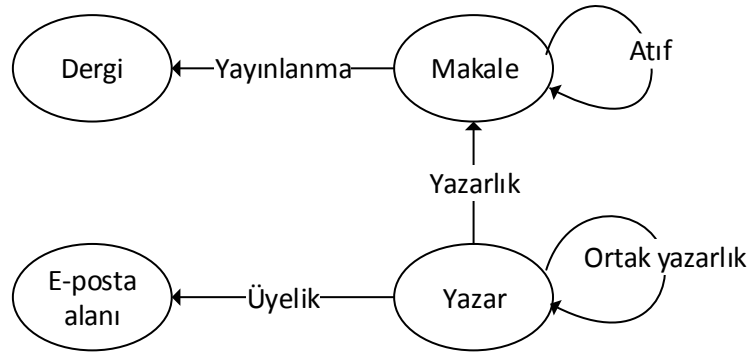
- **Standart bir deęerlendirme ölçütünün ve standart kıyaslama verilerinin olmaması:** Önerilen bağlantı tahmini yöntemleri için kıyaslama yapılabilecek standart bir deęerlendirme ölçütü ve standart veri kümeleri yoktur. Bu nedenle önerilen yöntemlerin doğru deęerlendirilmesi eksik kalmaktadır. Bir yöntemin performansı bazı veri kümelerinde iyi olabilirken, dięer veri kümelerinde yeterli olmayabilir. Yöntemlerin doğru deęerlendirilmesi için standart bir deęerlendirme ölçütüne ve testlerin yapılacağı standart veri kümelerine ihtiyaç vardır.



3. TEZ ÇALIŞMASINDA KULLANILAN AĞ VERİ KÜMELERİ VE DEĞERLENDİRME ÖLÇÜTLERİ

3.1. Kullanılan Ağ Verileri

- **Yüksek Enerji Fiziği Teorisi (High Energy Physics Theory, HEP-Th) Veri Kümesi:** HEP-Th [90] ağ veri kümesi, Massachusetts Amherst Üniversitesi'nin Bilgi Keşfi Laboratuvarı (Knowledge Discovery Laboratory, KDD) tarafından sağlanmış bilimsel çalışmalarla ilgili bibliyografik bilgilerin tutulduğu bir veri kümesidir. Bu veri kümesi KDD 2003 yarışması için arXiv arşivinden ve Stanford Lineer Hızlandırıcı Merkezi (Stanford Linear Accelerator Center, SPIRES-HEP) veri tabanından oluşturulmuştur. Bu ağ veri kümesi fizik biliminin teorik yüksek enerji alanında yapılmış bilimsel çalışmalarla ilgili 42.000'den fazla düğüm ve bu düğümler arasında 500.000'den fazla bağlantı olmak üzere zengin bibliyografik bilgiler içerir. Ağın düğümleri 1992-2003 yılları arasında yayınlanmış bilimsel makalelerle ilgili 448 dergi, 9200 yazar, 29.555 makale ve 3116 kurumun e-posta alan bilgilerinden oluşur. Ağın bağlantıları ise 87.794 ortak yazarlık, 352.807 atıf, 58.515 yazarlık, 20.826 yayın ve 12.487 e-posta kurum üyelik bilgilerinden oluşur. HEP-Th ağının düğüm ve bağlantı şeması Şekil 3.1'de verilir.



Şekil 3.1. HEP-Th ağı şeması

- **Bilgisayar Bilimi Bibliyografya (Computer Science Bibliography, DBLP) Veri Kümesi:** DBLP ağ veri kümesi [91], DBLP Bilgisayar Bilimi Bibliyografyası (Computer Science Bibliography, [92]) veri tabanından çıkarılan bilgisayar bilimi yayınları hakkında bilgiler içermektedir. DBLP ağının düğümleri makale, konferans,

tez, kitap, dergi, yazar ve web sayfası adres, gibi nesnelerden; ağı bağlantıları ise yazarlık, yayınlandığı dergi yayınlandığı konferans, atıf, editörlük gibi ilişkilerden oluşur. DBLP Bilgisayar Bilimi Bibliyografyası [92] veri tabanında 1,7 milyondan fazla yazar, 3,3 milyon makale, 32.000 dergi ve 31.000 konferans bilgisi bulunur. Massachusetts Amherst Üniversitesi'nin Bilgi Keşfi Laboratuvarı tarafından bu veri tabanından DBLP ağ veri kümesi [91] oluşturulmuştur. Bu ağ veri kümesi 1936-2006 yılları arasındaki 270.457 makale, 456.385 yazar ve 112.303 atıf bilgisinden oluşur.

- **Aminer:** Aminer atıf ağı veri kümesi DBLP [92] , Bilişim Makineleri Derneği [93] (Association for Computing Machinery, ACM) ve diğer kütüphanelerden elde edilen bilim dünyasıyla ilgili bibliyografik verilerden oluşturulmuştur. Veri kümesi; makalelerin özet, yazar, yıl, yayınlandığı yer ve başlık bilgileri gibi zengin içeriğe sahiptir. Bu veri kümesinde 1900-2011 yılları arasında yayınlanmış 1.397.240 makale ve 3.021.489 atıf bilgisi bulunur.

3.2. Değerlendirme Ölçütleri

Gerçek karmaşık ağların büyük çoğunluğu çok seyrek, yani aralarında bağlantı olan düğüm çiftlerinin sayısı aralarında bağlantı olmayan düğüm çiftlerinin sayısına oranla çok azdır. Bu durum, sınıf dengesizliği problemine yol açar ve önerilen bir bağlantı tahmini yönteminin gerçek performansının doğru bir şekilde değerlendirilmesini zorlaştırır. Bağlantı tahmini için standart bir değerlendirme ölçütü yoktur.

Alıcı İşlem Karakteristikleri (Receiver Operating Characteristic, ROC) eğrisi bağlantı tahmini için yaygın kullanılan ölçütlerden biridir. ROC eğrisi sınıflandırıcıların performanslarının karşılaştırılması ve görselleştirilmesi için kullanılan bir tekniktir. İlk olarak radar sinyallerinde dost düşman ayrımının tespiti için kullanılmıştır. Daha sonra tıp, biyoloji, psikoloji, makine öğrenmesi gibi çeşitli alanlarda oldukça yaygın bir kullanımı olmuştur. Makine öğrenmesi çalışmalarında verilerde sınıf dengesizliği olduğu durumlarda performansların doğru değerlendirilmesi için ROC eğrisi kullanışlıdır. Çünkü test verilerindeki sınıf dağılımları değişse bile ROC eğrileri değişmez. ROC eğrisi tüm sınıflandırma eşiklerinde yanlış pozitif orana (FPR) göre doğru pozitif oranı (TPR) gösteren bir eşik eğrisidir. Sınıflandırıcının bir örnek için ürettiği skor değeri eşik değerinden yüksekse örneğin sınıfı pozitif, değilse negatif olarak belirlenir. TPR, doğru sınıflandırılan

pozitif örneklerin sayısının tüm pozitif örneklerin sayısına oranıdır. FPR, yanlış sınıflandırılan negatif örneklerin sayısının toplam negatif örnek sayısına oranıdır.

Bağlantı tahmini için yaygın olarak kullanılan değerlendirme ölçütlerinden biri de ROC Eğrisi Altındaki Alan (AUROC). AUROC, ROC eğrisi altındaki alanın değeridir. AUROC değerleri 0 ile 1 arasındadır, 0,5'ten büyük değerler tahmin yönteminin ikili sınıflar için rastgele tahminden daha iyi olduğunu gösterir. AUROC; bağlantı tahmini açısından, rastgele seçilmiş gerçekte var olan bir bağlantının tahmin puanının rastgele seçilmiş gerçekte var olmayan bir bağlantının tahmin puanından yüksek olma ihtimalidir. AUROC denklem 3.1'e göre hesaplanır.

$$AUC = \frac{n' + 0.5n''}{n} \quad (3.1)$$

n , kaç tane bağımsız karşılaştırmanın yapıldığını; n' gerçekte var olan bağlantıların tahmin puanının gerçekte var olmayan bağlantıların tahmin puanından kaç kez daha yüksek olduğunu ve n'' ; gerçekte var olan bağlantıların tahmin puanının var olmayan bağlantıların tahmin puanına kaç kez eşit olduğunu gösterir.

4. BİLİM İNSANLARININ ATIF SAYISINI TAHMİN ETMEK İÇİN YENİ BİR BAĞLANTI TAHMİNİ YAKLAŞIMI

Bilim dünyasında bilimsel etkilerin tahmini problemi ile ilgili çalışmalar çoğunlukla bilimsel makalelerin alacağı atıf sayısının tahmin edilmesine odaklanmıştır. Literatürde bilim insanlarının bireysel olarak etkilerini/performanslarını tahmin etmek için çok az sayıda araştırma yapılmıştır. Bilim insanlarının bireysel etkilerini tahmin etmek, yapılacak bilimsel araştırmalar ve organizasyonlar için önemlidir. Bu açıdan, bilim insanlarının atıf sayılarını tahmin etmek için bu bölümde yeni bir yöntem önerilmiştir. Makalelerin atıf sayısını tahmin eden çalışmaların birçoğunda makalenin başlığı, özeti, anahtar kelimeleri; bilim insanlarının görev yaptığı enstitü, yayıncıların etki faktörü vb. gibi doküman tabanlı nitelikler kullanılmıştır. Ancak makaleler arasındaki atıfların oluşturduğu karmaşık ağların topolojik niteliklerinden pek yararlanılmamıştır. Hâlbuki atıf ağları hem makalelerin atıf sayısı tahmini hem de bilim insanlarının atıf sayısı tahmini için değerli nitelikler barındırmaktadırlar. Bu nedenle bu çalışmada bilim insanlarının atıf sayısı tahmini problemi yönlü, ağırlıklı ve zamansal atıf ağlarında bir bağlantı tahmini problemi olarak formüle edilmiştir.

Önerilen yaklaşım sadece bağlantıları değil aynı zamanda bağlantıların ağırlıklarını da tahmin edebilmektedir. Bu amaçla, ilk adım olarak atıf ağının zamansal gelişimi boyunca artan/azalan eğilimlerini dikkate alan zamansal bir bağlantı tahmini ölçütü önerilmiştir. Önerilen bağlantı tahmini ölçütü atıf sayısı tahmini için eğitici öğrenme algoritmalarında kullanılmıştır. Deney 1’de iki atıf ağında eğitici bir bağlantı tahmini yöntemi ile atıf sayısı tahmini yapılmıştır. Hem bağlantı hem de ağırlık tahminin yönlü, ağırlıklı ve zamansal ağlarda ilk kez ele alındığı göz önünde bulundurulduğunda Deney 1’deki testlerde başarılı sonuçlar elde edilmiştir. Deney 2’de, önerilen bağlantı tahmini ölçütü ve geleneksel bir bağlantı tahmini ölçütü olan AA ölçütünün performansı karmaşık ağlarda yeni bağlantı tahmini kapsamında karşılaştırılmıştır. Karşılaştırma testlerinin tümünde önerilen bağlantı tahmini yönteminin daha iyi sonuçlar elde ettiği görülmüştür.

Alıntı sayısı, bilim insanlarının ve makalelerinin etkisini değerlendirmek için kullanılan geleneksel bir ölçüttür. Bir bilim insanının kalitesini belirlemek için yaygın olarak kullanılan h-index [94] ölçütü de makalelerin atıf sayısı ile hesaplanmaktadır. Bilim dünyasında bilimsel etki tahmini problemini ele alan birçok çalışmada [95-101] makalelerin atıf sayı tahmin edilmeye çalışılmıştır. Bu çalışmalarda genellikle makalelerin başlığı, özeti, anahtar kelimeleri, bilim insanlarının görev yaptığı enstitü bilgileri, yayıncının etki faktörü gibi

doküman tabanlı nitelikler kullanarak regresyon modelleri uygulanmıştır. Bunların çoğunda atıf ağlarının topolojik nitelikleri kullanılmamıştır. Bilimsel makalelerdeki referanslar, bilimsel etkileri araştırmak için kullanılabilecek karmaşık atıf ağlarını oluşturmaktadırlar. Çok az sayıdaki çalışmada atıf ağlarının topolojik yapılarından faydalanılmıştır. Pobiedina ve Ichise [95] tarafından makalelerin atıf sayısı tahmini için atıf ağlarındaki topolojik nitelikleri kullanılmıştır. Atıf ağlarında sıkça gösterilen topolojik motiflere dayalı yeni bir nitelik önerilmiş ve bu nitelik sınıflandırma ve regresyon modellerinde makalelerin atıf sayısı tahmini için ayrı ayrı test edilmiştir.

Literatürdeki pek çok bilimsel makale ortak yazarlıdır. Bu nedenle makaleler bilim insanlarının potansiyel etkilerini tek tek değerlendirmek için yetersiz kalabilirler. Bilim insanları farklı konularda çalışabilirler. Bir bilim insanının genel performansı iyi olmasına rağmen belirli bir konudaki yetkinliği iyi bir seviyede olmayabilir. Ayrıca bir bilim insanının performansı zaman içinde değişebilir. Geçmişte bir bilim insanı çok sayıda atıf alırken, yakın geçmişteki bilimsel etkinliği azalabilir. Bilim insanlarının bireysel atıf sayısı tahmini; bilimsel çalışmalara desteklerin/bursların verilmesi, proje başvuruların kabul edilmesi, hakem/bilirkişi belirlenmesi vb. konularında karar vermeye yardımcı bir kıstas olarak kullanılabilir. Bu gerekçelerden dolayı, bilim insanları arasındaki karmaşık atıf ilişkilerini kullanan eğiticili bir atıf sayısı tahmini yaklaşımı bu bölümde önerilmektedir.

Sadece birkaç çalışmada bilim insanlarının atıf sayısı tahmini problemi ele alınmıştır. Dong vd. [102] tarafından bilim insanlarının etki tahmini ve makalelerinin etki tahmini olmak üzere iki problem ele alınmıştır. Bu çalışmanın [102] temel amacı, yeni yayınlanmış bir makalenin o makalenin birinci yazarının gelecekteki h-indeksine katkıda bulunup bulunmayacağını tahmin etmektir. Bu amaca yönelik olarak, birinci yazarın gelecekteki h-indeksi tahmin edilmeye çalışılmıştır. Temel amacı gerçekleştirmek için birinci yazarın tahmin edilen h-indeksi kullanılmıştır. Önerilen tahmin edici modellerde yazar, makale, dergi, sosyal, referans ve zaman tabanlı nitelikler incelenmiştir. Makale konusu ve makalenin yayınlanacağı derginin tahmin yöntemindeki en önemli nitelikler olduğu tespit edilmiştir. Revesz [103] tarafından bilim insanlarının atıf eğrilerini tahmin etmeye yönelik bir yaklaşım önerilmiştir. Makalelerin derecelerini ve yayın tarihlerini atıf sayısına yaklaştıran uzaysal ve zamansal bir matematiksel fonksiyon ile atıf eğrisi tahmin edilmiştir. Bu çalışmaların ikisinde de atıf ağlarındaki topolojik nitelikler kullanılmamıştır.

Bu bölümde ilgilenilen bilim insanlarının atıf sayısı tahmini problemi yönlü, ağırlıklı ve zamansal atıf ağlarında bir bağlantı tahmini problemi olarak formüle edilmiştir. Problem,

bağlantılarla beraber ağırlığın da tahminini içerdiğinden daha fazla çaba gerektirmektedir. Bu çalışma bilindiği kadarıyla yönlü, ağırlıklı ve zamansal ağlarda bağlantıların ağırlıklarıyla birlikte tahmin edildiği ilk çalışmadır. Bu işi gerçekleştirmek için öncelikle zamansal bir bağlantı tahmini ölçütü önerilmiş, daha sonra önerilen bu ölçüt eğitici bağlantı tahmini algoritmalarında kullanılmıştır. Statik ağlar için önerilen bağlantı tahmini yöntemlerde düğüm çiftleri arasında görülen eski, yeni veya tekrarlanan bağlantılar ayırt edilemez. Ancak sosyal ağlarda bağlantıların eklenmesi/çıkarılması işlemleri sıklıkla gözlemlenir. Dinamik ağlarda yapılan bağlantı tahmini çalışmalarında [104-107] zamana dayalı algoritmaların bağlantı tahmin performansını geliştirdiği görülmektedir.

Literatürdeki bağlantı tahmini ölçütleri yerel ve küresel ölçütler olarak kategorize edilebilir. Yerel ölçütler Adamic/Adar [47] ve CAR indeksi [108] gibi düğümlerin komşuları veya komşularının komşularıyla hesaplanmaktadır. Yerel ölçütlerde hesaplama verimliliği vardır, ancak ağdaki karmaşık yapılar sınırlı düzeyde ele alınabilir. Küresel ölçütlerde, düğümler arasındaki bağlantılar üzerinde oluşan yollar kullanılır. Katz [56], Rooted PageRank [109] ve SimRank [60] küresel tabanlı bağlantı tahmini ölçütlerine birer örnektir. Küresel bağlantı tahmini ölçütlerinde ağlardaki karmaşık topolojik yapılar etkili bir şekilde ele alınabilir, ancak hesaplama maliyetleri, yerel tabanlı ölçütlerden daha yüksektir. Bu çalışmada karmaşık ağlardaki hem yerel hem de küresel topolojik nitelikleri ele alabilen bir bağlantı tahmini ölçütü önerilmiştir.

Bu bölümde yapılan çalışmanın ana katkıları şöyle özetlenebilir:

- Bilim insanlarının atıf sayısı tahmini problemini ele alan çok az sayıdaki çalışmalara katkıda bulunulmuştur.
- Karmaşık ağlarda hem yerel hem de küresel topolojik yapılardan yararlanan yeni bir zamansal bağlantı tahmini ölçütü önerilmiştir.
- Bilim insanlarının atıf sayısı tahmini problemi atıf ağlarında bir bağlantı tahmini problemi olarak formüle edilmiştir. Önerilen bağlantı tahmini yaklaşımında bağlantılar ağırlıklarıyla birlikte tahmin edilmiştir. Bu açıdan yapılan çalışma yönlü, ağırlıklı ve zamansal ağlarda bağlantıların ağırlıklarıyla beraber tahmin edildiği ilk çalışmadır.

4.1. Zamansal Bağlantı Tahmini Ölçütü

Önerilen zamansal bağlantı tahmini ölçütünde ağdaki v_s ve v_t düğüm çiftlerinin yakınlığını hesaplanır. Ölçütün ürettiği yüksek skor v_s düğümünden v_t düğümüne bağlantı

oluşma olasılığının yüksek olduğunu gösterir. Önerilen ölçüt zaman aralıkları boyunca hesaplanır. Verilen bir ağ $G_{t,t'}(V, E)$ belirlenen zaman aralıklarına göre alt ağlara bölünür. Zaman aralıklarına göre ayrılmış alt ağlar ve f . alt ağ sırasıyla şöyle tanımlanmıştır:

$$G_{t,t'}(V, E) = \left\{ \begin{array}{c} G_{t+0w, t+1w-1}(V, E), \\ G_{t+1w, t+2w-1}(V, E), \\ \vdots \\ G_{t+(n-1)w, t+nw-1}(V, E) \end{array} \right\} \quad (4.1)$$

$$G_f = G_{t+(f-1)w, t+fw-1}(V, E) \quad (4.2)$$

Denklemlerde verilen ağlar düğümler arasında belirtilen zaman aralıklarında oluşan bağlantıları içermektedirler. Önerilen bağlantı tahmini ölçütü belirlenen zaman aralıklarının geçişleri sırasında oluşan zamansal olaylar üzerinden hesaplanır. Kullanılan zamansal olaylar şöyle tanımlanır.

4.1.1. Zamansal Olaylar

Soares ve Prudencio [46] tarafından önerilen zamansal olaylar bu çalışmada ağırlıklı, yönlü ve mesafeli olarak geliştirilmiştir. Önerilen zamansal olaylar doğrudan ve dolaylı olmak üzere iki tiptir. Doğrudan zamansal olaylar yerel topolojik yapılarda gerçekleşirken dolaylı zamansal olaylar küresel topolojik yapılarda gerçekleşmektedir.

4.1.1.1. Doğrudan Zamansal Olaylar

Doğrudan bir zamansal olay, $f - 1$ zaman aralığından f zaman aralığına geçiş sırasında v_s , v_t düğümleri ve bunların ortak komşuları arasında bir bağlantının silinmesi/oluşması, güçlenmesi/zayıflaması veya ağırlığının korunması olaylarını içermektedir. Tutucu, yenilikçi ve geriletici olmak üzere üç doğrudan olay vardır.

1) Tutucu Olay: Tutucu olayda ardışık zaman çerçeveleri geçişinde v_s ve v_t düğümleri arasındaki bağlantı tekrar eder. G_{f-1} ağında v_s ile v_t arasındaki bağlantının ağırlığı G_f ağında artabilir, azalabilir veya değişmeyebilir. Denklem 4.3'te verilen tutucu olay formülünde bu üç durum değişik oranlarda skora pozitif katkı sağlayacak şekilde yansıtılır. Çünkü bağlantının ardışık zaman çerçevelerinde farklı ya da aynı ağırlıkla devam etmesi v_s ile v_t

düğümüleri arasındaki yakınlığın korunduğunu gösterir. Bu açıdan Tablo 4.1’de tutucu olay için verilen 4.4, 4.5 ve 4.6 denklemlerindeki c katsayısı pozitif sabit bir değerdir. Bağlantının ağırlığı G_{f-1} ve G_f ağlarında aynı olduğunda skor $w(u, v, f)$ değeri ile c sabitinin çarpımı şeklinde hesaplanır. Bu durumda skor değerinin pozitif olarak artırılması korunan bağlantının ağırlığı ile orantılıdır. Bağlantının ağırlığı $f - 1$ ’den f ’ye geçildiğinde zayıflamışsa, skor $w(u, v, f)$ ’ye göre yükseltilir, diğer yandan ağırlıktaki azalma oranına göre düşürülür. Son durum ise zaman aralıkları geçişlerinde bağlantının ağırlığının artması durumudur. Bu durumda skor $w(u, v, f - 1)$ ile orantılı olarak artırılır, ayrıca geçişteki artış oranına göre skor değeri pozitif olarak yükseltilir.

2) Yenilikçi Olay: Bu olay, v_s ile v_t düğümlerinin G_{f-1} ağında birbirlerinden uzaktayken G_f ağında birbirlerine yaklaşmakta olduğunu gösterir. Yenilikçi olay denklem 4.7’de $w(u, v, f)$ ’nin pozitif i katsayısı ile çarpılmasıyla hesaplanır.

3) Geriletici Olay: G_{f-1} ağında v_s ile v_t arasında görülen bağlantının G_f ’de kaybolduğu olaydır. Bu olay, v_s ile v_t arasındaki ilginin azaldığını gösterir. Geriletici olay denklem 4.8’de negatif c sabitinin $w(u, v, f - 1)$ ile çarpılmasıyla hesaplanır.

Tablo 4.1. Doğrudan Zamansal Olay Denklemleri

Zamansal Olaylar	Denklemler	
Tutucu Olay	$C(v_s, v_t, f) = \begin{cases} C_d, & w(v_s, v_t, f - 1) > 0 \wedge w(v_s, v_t, f) > 0 \wedge w(v_s, v_t, f - 1) > w(v_s, v_t, f) \\ C_i, & w(v_s, v_t, f - 1) > 0 \wedge w(v_s, v_t, f) > 0 \wedge w(v_s, v_t, f - 1) < w(v_s, v_t, f) \\ C_u, & w(v_s, v_t, f - 1) > 0 \wedge w(v_s, v_t, f) > 0 \wedge w(v_s, v_t, f - 1) = w(v_s, v_t, f) \\ 0, & \text{değilse} \end{cases}$	(4.3)
	$C_d(v_s, v_t, f) = c * \left[w(v_s, v_t, f) - \left((w(v_s, v_t, f - 1) - w(v_s, v_t, f)) / w(v_s, v_t, f - 1) \right) \right]$	(4.4)
	$C_i(v_s, v_t, f) = c * \left[w(v_s, v_t, f - 1) + (w(v_s, v_t, f) / w(v_s, v_t, f - 1)) \right]$	(4.5)
	$C_u(v_s, v_t, f) = c * w(v_s, v_t, f)$	(4.6)
Yenilikçi Olay	$N(v_s, v_t, f) = \begin{cases} i * w(v_s, v_t, f), & w(v_s, v_t, f - 1) = 0 \wedge w(v_s, v_t, f) > 0 \\ 0, & \text{değilse} \end{cases}$	(4.7)
Gerileyici Olay	$D(v_s, v_t, f) = \begin{cases} r * w(v_s, v_t, f - 1), & w(v_s, v_t, f - 1) > 0 \wedge w(v_s, v_t, f) = 0 \\ 0, & \text{değilse} \end{cases}$	(4.8)

Tablo 4.2. Dolaylı Zamansal Olaylar

Denklemler	
$I(v_s, v_t, f) = \begin{cases} I_d, & 1/d(v_s, v_t, f-1) > 1/d(v_s, v_t, f) \\ I_i, & 1/d(v_s, v_t, f-1) < 1/d(v_s, v_t, f) \\ I_u, & 1/d(v_s, v_t, f-1) = 1/d(v_s, v_t, f) \end{cases}$	(4.9)
$I_d(v_s, v_t, f) = c * \left[1/d(v_s, v_t, f) - \left((1/d(v_s, v_t, f-1) - 1/d(v_s, v_t, f)) / 1/d(v_s, v_t, f-1) \right) \right]$	(4.10)
$I_i(v_s, v_t, f) = c * \left[1/d(v_s, v_t, f-1) + (1/d(v_s, v_t, f) / 1/d(v_s, v_t, f-1)) \right]$	(4.11)
$I_u(v_s, v_t, f) = c * 1/d(v_s, v_t, f)$	(4.12)

4.1.1.2. Dolaylı Zamansal Olaylar

Dolaylı zamansal olaylar, v_s ile v_t arasındaki mesafenin G_{f-1} 'den G_f 'ye geçiş esnasında artması, azalması ya da değişmemesi durumlarını kapsamaktadır. Mesafe v_s ile v_t düğümlerini ağda bağlantılar üzerinden birleştiren en kısa yolun uzunluğudur. Dolaylı zamansal olayların denklemleri Tablo 4.2'de verilmiştir. v_s ve v_t düğümleri arasındaki mesafenin uzaması, bu düğümler arasında gelecekte bağlantı oluşma ihtimalinin düşük olacağı anlamına gelir. Dolaylı zamansal olayın hesaplanması tutucu olaya benzemektedir. Tutucu olayda sadece v_s , v_t düğümleri ve bunların ortak komşuları arasındaki bağlantılar dikkate alınırken dolaylı zamansal olayda v_s ve v_t düğümlerinin karmaşık bir ağda birbirlerine ne kadar yakın olduğu incelenir. Bu nedenle dolaylı zamansal olaylar karmaşık ağların zaman içerisindeki gelişimini anlamada önemlidir.

4.1.1.3. Önerilen Zamansal Bağlantı Tahmini Ölçütünün Hesaplanması

Önerilen zamansal bağlantı tahmini ölçütünün hesaplanması denklem 4.14 ve 4.15'te sırasıyla verilen birincil ve ikincil ölçütlerden oluşur. Her iki ölçüt doğrudan zamansal olayları kullanır. Tutucu, yenilikçi ve gerileyici olayların skoru, denklem 4.13'te $f-1$ 'den

f 'ye geçiş sırasında toplanır. Denklemler 4.14 ve 4.16'da yakın zamanda gerçekleşen olayların skor değerini yükseltmek için zamansal bir logaritmik fonksiyon kullanılır.

Denklem 4.14'te verilen birincil ölçütte v_s ve v_t düğümleri arasında gerçekleşen olaylar hesaplanır. Birincil ölçütte doğrudan ve dolaylı zamansal olaylar kullanılır. $f - 1$ 'den f 'ye geçiş sırasında doğrudan bir zamansal olay meydana gelmişse doğrudan zamansal olay skoru hesaplanır, aksi takdirde dolaylı zamansal olay skoru hesaplanır. Bu işlem tüm zaman aralıkları geçişleri için uygulanır.

Denklem 4.15'te verilen ikincil ölçütte ise v_s ve v_t düğümlerinden biri ile bu düğümlerin ortak komşularından biri arasında gerçekleşen olaylar hesaplanır. v_s ve v_t düğümlerine yakın olan düğümler üzerinde gerçekleşen olaylar v_s ve v_t düğümleri arasında bağlantı oluşmasında daha belirleyicidir. Bu nedenle v_s ve v_t 'nin tüm ortak komşuları üzerinde zaman aralıkları geçişleri boyunca gerçekleşen olaylar hesaplanır.

Denklem 4.17 önerilen zamansal bağlantı tahmini ölçütünün son şeklidir. v_s ve v_t düğümleri arasındaki olaylar (birincil olaylar) bunların ortak komşuları ile aralarında gerçekleşen olaylardan (ikinci olaylar) daha önemli olduğundan denklem 4.17'de ikincil olayların skor değeri $[0,1]$ aralığında belirlenen sabit bir değer olan α ile azaltılır. Önerilen yöntemde birincil ölçüt sayesinde küresel düzeyde karmaşık yapılardan yararlanılırken ikincil ölçütlerle karmaşık ağlardaki yerel yapılar göz önünde bulundurulur.

$$E(v_s, v_t, f) = C(v_s, v_t, f) + N(v_s, v_t, f) + D(v_s, v_t, f) \quad (4.13)$$

$$P(v_s, v_t) =$$

$$\sum_{f=2}^n \begin{cases} \log(f+1) * E(v_s, v_t, f), & \text{doğrudan bir zamansal olay varsa} \\ \log(f+1) * I(v_s, v_t, f), & \text{yoksa dolaylı zamansal olay hesaplanır} \end{cases} \quad (4.14)$$

$$S(v_s, v_t) = \sum_{y \in \Gamma(v_s) \cap \Gamma(v_t)} \frac{L(v_s, y) + L(y, v_s)}{L(v_s, y) + L(y, v_t)} \quad (4.15)$$

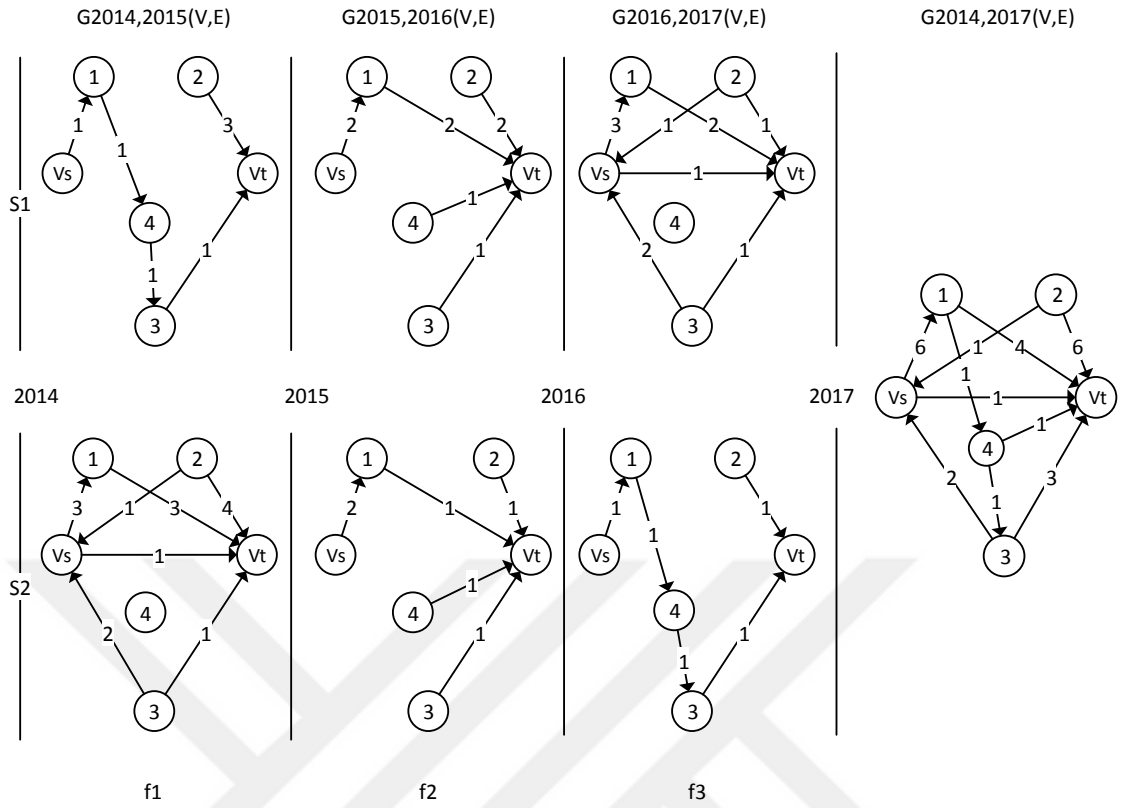
$$L(v_s, v_t) = \sum_{f=2}^n \log(f+1) * E(v_s, v_t, f) \quad (4.16)$$

$$Skor(v_s, v_t) = P(v_s, v_t) + \alpha * S(v_s, v_t) \quad (4.17)$$

Önerilen bağlantı tahmini ölçütünün hesaplanmasının gösterilmesi için Şekil 4.1'de örnek bir $G_{2014,2017}(V,E)$ ağı bulunur. Farklı zamansal eğilimlerin etkisini göstermek üzere $G_{2014,2017}(V,E)$ ağı S1 ve S2 olmak üzere iki ayrı zamansal senaryo şeklinde gösterilir. S1 ve S2 genişliği bir olan f1, f2 ve f3 zaman aralıklarına bölünür. Ağın zaman aralığı gösterimlerinde her bir zaman aralığında sadece mevcut zaman çerçevesinde oluşan bağlantılar gösterilir. S1 ve S2'deki (v_s, v_t) düğüm çiftinin önerilen bağlantı tahmini ölçütü skorları Tablo 4.3'te verilir. c , r ve i parametrelerine sırasıyla 0,5, -0,5 ve 1 değerleri verilir.

S1 ve S2'deki senaryolarda farklı eğilimler varken statik bağlantı tahmini ölçütlerinde (v_s, v_t) düğüm çifti için aynı skor elde edilir. Ancak (v_s, v_t) düğüm çifti ve bu düğümlerin ortak komşuları arasında meydana gelen zamansal olaylar S1 senaryosunda yükselme eğilimi gösterirlerken S2 senaryosunda düşüş eğilimi gösterirler. Örneğin, S1 senaryosunun f3 çerçevesinde yenilikçi birincil bir olay gerçekleşirken S2 senaryosunun f2 çerçevesinde geriletici birincil bir olay gerçekleşir. Önerilen zamansal bağlantı tahmini ölçütünde yakın zamanlardaki olaylar daha fazla ağırlıklandırılır. Ayrıca ağın gelişimi boyunca v_s ve v_t düğümleri arasındaki mesafenin artıp azalması birincil ölçütte hesaplamaya katılır. S1 senaryosundaki zaman çerçevelerinde v_s ve v_t düğümleri arasındaki mesafe değerleri sırasıyla 4, 2 ve 1 iken S2 senaryosunda aynı düğüm çiftleri için mesafe değerleri 1, 2 ve 4'tür. Bu nedenle, S1 senaryosundaki (v_s, v_t) düğüm çiftinin birincil skoru (2,61), S2 senaryosundaki birincil skorundan (-0,71) daha büyüktür.

Farklı eğilimler ikincil olaylarda da vardır. S1 ve S2 senaryolarındaki $(1, v_s)$ ve $(1, v_t)$ düğüm çiftleri arasındaki bağlantıların 2014 ile 2017 yılları arasındaki toplam ağırlığı aynı iken söz konusu düğüm çiftleri S1 senaryosunda yükselme eğilimi S2 senaryosunda düşüş eğilimi gösterir. 1 nolu düğümün v_s ve v_t düğümleri ile S1 ve S2'de gerçekleştirdiği ikincil olayların skorları sırasıyla 7,65 ve 1,43'tür. Görüldüğü gibi önerilen zamansal bağlantı tahmini ölçütüyle artan ve azalan eğilimler skora yansıtılır. Benzer şekilde $(2, v_t)$ düğüm çifti arasındaki geçici olaylar S1 ve S2'de farklı eğilimler gösterir. Bu düğüm çifti arasındaki bağlantının ağırlığı, hem S1 hem de S2'de f1'den f3'e kadar azalır. Ancak S1'deki geçişler sırasında bağlantının ağırlığı kademeli olarak zayıflarken, S2'de aniden zayıflar. Önerilen ölçüt ile bu durumlar da birbirinde farklı şekilde değerlendirilir, S1 ve S2'de $(2, v_t)$ düğüm çifti arasında meydana gelen olayların skorları sırasıyla 1,26 ve 0,83'tür.



Şekil 4.1. Örnek bir ağ

Tablo 4.3. Örnek ağın skor değerleri

		(v_s, v_t) düğüm çiftinin S1'deki skorları		(v_s, v_t) düğüm çiftinin S2'deki skorları	
		$f1 \rightarrow f2$	$f2 \rightarrow f3$	$f1 \rightarrow f2$	$f2 \rightarrow f3$
Birincil Skorlar	Doğrudan	Zamansal olaylar: -0	Zamansal olaylar: N 1,38	Zamansal olaylar: D -0,54	Zamansal olaylar: -0
	Dolaylı	Zamansal olaylar: I 1,23	Zamansal olaylar:-0	Zamansal olaylar:-0	Zamansal olaylar: D -0,17
İkincil Skorlar	Düğüm1	Zamansal olaylar: C,N 3,84	Zamansal olaylar: C 3,81	Zamansal olaylar: C 1,09	Zamansal olaylar: C, D -0,34
	Düğüm2	Zamansal olaylar: C 0,91	Zamansal olaylar: C, N 1,73	Zamansal olaylar: C, D -0,41	Zamansal olaylar: C 0,69
	Düğüm3	Zamansal olaylar: C 0,54	Zamansal olaylar: C,N 3,46	Zamansal olaylar: C,D -0,54	Zamansal olaylar: C 0,69
Toplam birincil skor		2,61		-0,71	
Toplam ikincil skor		14,29		1,18	
Skor(v_s, v_t)		$2,61 + \alpha * 14,29$		$-0,71 + \alpha * 1,18$	

Tablo 4.4. Sınıf kategorileri

Sınıf kategori adı	Sınıf adı	Sınıfların atıf sayısı aralıkları
C4	sınıf0	0
	sınıf1	[1,2]
	sınıf2	[3,5]
	sınıf3	>5
C5	sınıf0	0
	sınıf1	1
	sınıf2	[2,3]
	sınıf3	[4,6]
	sınıf4	>6

Tablo 4.5. Kullanılan nitelik kümeleri

Nitelikler	Nitelik kategorileri		
	F1	F5	F8
$Skor(v_s, v_t)$	✓		
$P(v_s, v_t)$		✓	✓
$L(v_s, y)$		✓	✓
$L(y, v_s)$		✓	✓
$L(v_t, y)$		✓	✓
$L(y, v_t)$		✓	✓
v_t 'nin $G_{t,t'}(V, E)$ ağındaki toplam atıf sayısı			✓
v_t 'ye $G_{t,t'}(V, E)$ içinde kaç farklı bilim insanının atıfta bulunduğu			✓
v_t 'nin h_indeksi			✓

4.2. Eğitici Bağlantı Tahmini

Bağlantı tahmin problemini çözümü için sınıflandırma modelleri kullanılır. Tablo 4.4'te C4 ve C5 olmak üzere iki sınıf kategorisi belirlenir. Bu iki kategoride farklı atıf sayısı aralıklarında sırasıyla dört ve beş adet sınıf bulunur. Yapılacak sınıflandırma işi, $G_{t,t'}(V, E)$ 'de verilen bir bilim insanının $G_{t'+1,t'+w}(V, E)$ 'de alacağı atıf sayısına karşılık gelen sınıfının belirlenmesidir. Bilim insanının alacağı atıf sayısı tahmin edileceğinden

$G_{t'+1,t'+w}(V,E)$ 'deki yeni, yinelenen, kaybolan ve oluşmayan bağlantılar tahmin alanı dâhilindedir. Sınıflandırma işi WEKA isimli [110] veri madenciliği yazılımında gerçekleştirilir. Önerilen zamansal bağlantı tahmini ölçütü deneylerdeki sınıflandırıcılarda nitelik olarak kullanılır. Verilen bir (v_s, v_t) düğüm çifti için, önerilen zamansal ölçütün nihai skorunun ($Score(v_s, v_t)$) sadece kullanılması yerine birincil skor ($P(v_s, v_t)$) ve ikincil skorlar ($L(v_s, y)$, $L(y, v_s)$, $L(v_t, y)$, $L(y, v_t)$) ayrı birer nitelik olarak beş elemanlı bir nitelik kümesi de (F5) kullanılır. Tahminin doğruluğunun arttırılması için F8 nitelik kümesi F5 nitelik kümesinden ve belirlenen üç basit nitelikten oluşturulur. Bu üç basit nitelik $G_{t,t'}(V,E)$ 'de v_t 'nin toplam atıf sayısı, v_t 'ye kaç farklı bilim insanının atıfta bulunduğu sayısı ve v_t 'nin h-indeksidir. Bir bilim insanının h-indeks değeri, bu bilim insanının h tane makalesinin her birinin en az h tane atfının olduğunu ve geri kalan makalelerinin hiçbirinin h veya h'den fazla atfının olmadığını gösterir. Deneylerde kullanılan tüm nitelikler Tablo 4.5'te verilmiştir.

4.3. Deneyler

4.3.1. Deney Konfigürasyonları

Testler, K-En Yakın Komşu, Rastgele Orman, Lojistik Regresyon ve Çok Katmanlı Algılayıcılar olmak üzere dört sınıflandırıcıda gerçekleştirilir. Bu sınıflandırıcıların WEKA'daki karşılıkları sırasıyla IBk, RandomForest (RF), Lojistik (LR) ve MultilayerPerception (MLP)'dir. RF, MLP ve LR için varsayılan parametreler ayarlanır. IBk için komşuların sayısı (the number of neighbours, KNN) parametresi 5 olarak ayarlanır. Kullanılan sınıflandırıcıların tahmin hatası 10 kat çapraz doğrulama ile değerlendirilir. c , r , i ve α parametreleri sırasıyla 0,5, -0,5, 1 ve 0,05 olarak verilmiştir.

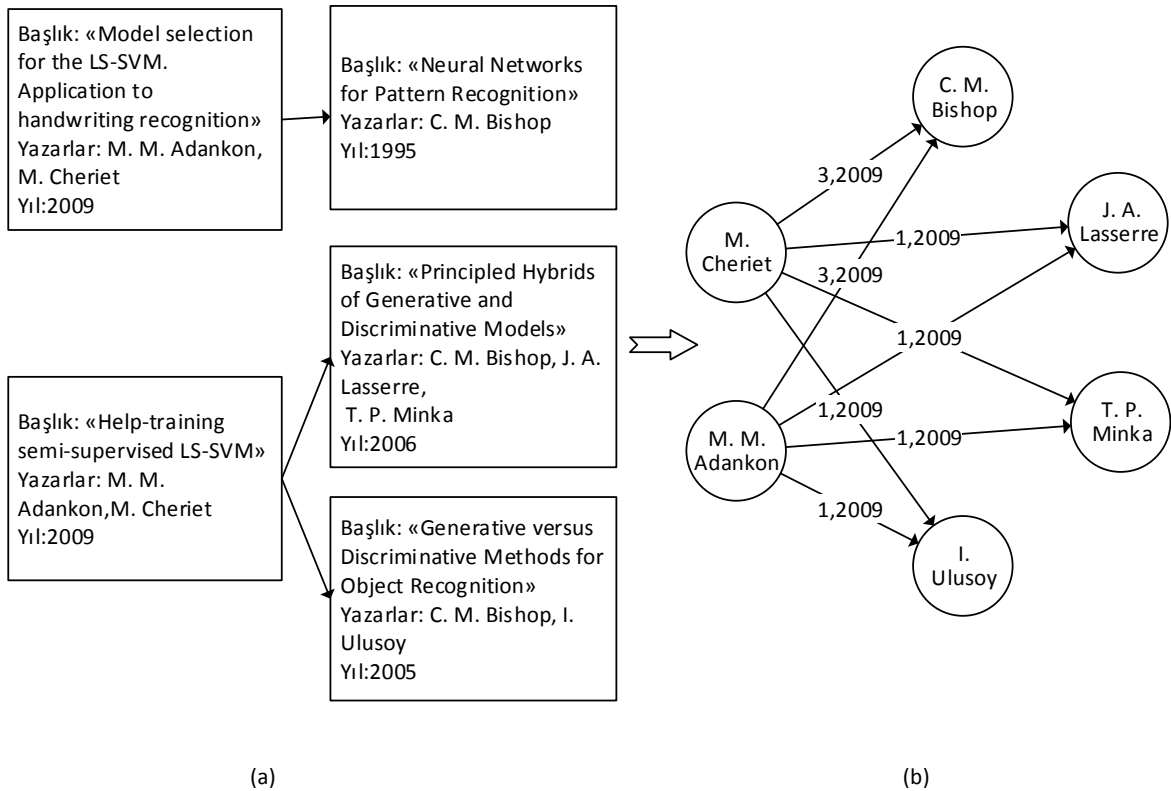
4.3.2. Test Veri Kümeleri ve Ağ Modeli

Önerilen yöntemin performansı HEP-Th ve Aminer veri kümelerinde test edilir. HEP-Th ve Aminer veri kümelerinden iki atıf ağı oluşturulur. Bu atıf ağlarında düğümler bilim insanlarını, yönlü bağlantılar bilim insanları arasındaki atıfları gösterir. Yönlü bağlantılar ağırlık ve yıl nitelik bilgilerini içerir. Kullanılan ağ modelinin örnek bir gösterimi Şekil

4.2'de verilir. Şekil 4.2 (a)'da birkaç makale ve bu makaleler arasındaki atıflar bulunur. Kullanılan ağ modelini gösteren Şekil 4.2 (b), Şekil 4.2 (a) 'nın yazar tabanlı atıf ağıdır. M. M. Adankon ve M. Cheriet'in yazar olduğu iki makale ve bu iki makaleden diğer üç makaleye yapılan üç atıf vardır. Atıf yapılan makalelerin hepsinde C. M. Bishop yazardır. Bu nedenle, yazar ağında M. M. Adankon ve M. Cheriet, C. M. Bishop'a 3 ağırlığında bir bağlantı ile bağlanır. Kullanılan ağ modeline göre, Aminer atıf ağı, 2002'den 2009'a kadar 12 milyondan fazla atıf ve 472.792 bilim insanı, HEP-Th atıf ağı ise 1992'den 2003'e kadar 735.710 alıntı ve 8380 bilim insanı verilerini içerir.

4.3.3. Deney Sonuçları

Deneyler iki kısımdan oluşur. Deney 1'de, temel problem olan bilim insanlarının atıf sayısı tahmini sonuçları verilir. Literatürdeki bağlantı tahmini yöntemlerinin çoğu karmaşık ağlarda yeni bağlantıların tahminine odaklanmıştır. Bu nedenle, Deney 2'de önerilen zamansal bağlantı tahmini ölçütü, yeni bağlantı tahmini kapsamında geleneksel bir bağlantı tahmini ölçütü olan AA [47] ölçütü ile karşılaştırılır.



Şekil 4.2. Ağ modeli için örnek bir gösterim

4.3.3.1. Deney 1: Bilim İnsanlarının Atıf Sayısı Tahmini

Sınıf dengesizliği problemini azaltmak için her bir sınıftaki örnek sayısı eşit olacak şekilde örnekleme yapılır. HEP-Th ve Aminer veri kümelerinin her biri için altı tane örneklendirilmiş veri kümesi oluşturulur. Eğitim veri setlerinde her yıl için en az 2 kez atıf alan bilim insanları içinden pozitif ve negatif sınıf örnekleri rastgele seçilir. Pozitif bir sınıf örneği iki bilim insanından oluşur ve test veri kümesinde bu iki bilim insanı arasında atıf bulunur, negatif sınıf örneğinde ise bilim insanları arasında atıf yoktur. Pozitif örnekler aralarında yeni veya tekrar atıf ilişkisi olan bilim insanlarından, negatif örnekler aralarındaki atıf ilişkisi tekrar etmeyen veya aralarında hiç atıf ilişkisi olmayan bilim insanlarından seçilir. Pozitif ve negatif örnekler eğitim veri ağında aralarında en fazla 5 birim uzaklık bulunan düğüm çiftlerinden seçilir. Negatif örnekler uzaklık durumuna bakılmaksızın rastgele seçilseydi gerçek ağlar oldukça seyrek yapıda olduğundan düğüm çiftlerinin çoğu büyük olasılıkla aralarında uzak mesafe olan düğüm çiftlerinde oluşacaktı. Pozitif örneklere benzeyen negatif örnekleri birbirinden ayırt etmek sınıflandırma işleminin en kritik noktasıdır. Aksi takdirde negatif olduğu kuvvetle muhtemel bir verinin negatif olduğunun tahmini, önerilen yöntemin etkin olduğunun anlaşılmasında yeterli değildir. Her bir sınıf için örnek sayısının eşit sayıda tutulması, pozitif ve negatif örneklerin en fazla 5 birim uzaklıkta bulunan ve her yıl en az 2 atıf alan düğüm çiftlerinden seçilmesiyle önerilen yöntemin performansının doğru bir şekilde değerlendirilebileceği eğitim verileri oluşturulur. Örnekleme yapılmış veri kümelerinin detayları Tablo 4.6'da verilir. Tablo 4.5'te her sınıf için atıf sayısı aralığının bulunduğu iki sınıf kategorisi listelenir. Problem, bağlantılar ve bağlantıların ağırlıklarının tahmini olduğundan pozitif örnekler Tablo 4.5'teki sınıflara karşılık gelen ağırlıklara göre örneklendirilir.

Örnekleme yapılmış veri kümelerinin isimleri veri kümelerinin özelliklerini gösteren bazı sembollere sahiptir. Örneğin Aminer için D3_C4_FW2_I atıf ağı veri kümesinde D3 veri kümesinin numarasını, C4 veri kümesinin C4 kategorisindeki sınıf örneklerini içerdiğini, FW2 zaman çerçeve genişliğinin iki olduğunu ve I, G_{f-1} 'in son yıl verilerinin G_f 'in ilk yıl verileriyle kesiştiğini gösterir. Bu veri kümesinin alt ağlarının zaman çerçeveleri kümesi {2002-2003, 2003-2004, 2004-2005, 2005-2006, 2006-2007} şeklindedir. Aminer'ın D1_C5_FW1 veri kümesinde, sınıfların örnek sayılarının eşit tutulabileceği ancak 11.560 örnek mevcuttur. Bu veri kümesinde her sınıfta 2312 örnek bulunur. Benzer şekilde HEP-Th ağında D1_C4_FW1 ve D1_C5_FW1 veri kümeleri için sırasıyla 9276 ve 7810 uygun

örnek bulunur. Bu örneklenmiş veri kümeleri sırasıyla her bir sınıf için 2319 ve 1562 örnek içerir. Aminer ve HEP-Th'nin diğer örnekendirilmiş veri kümelerindeki her bir sınıfa ait örnek sayısı 3000 olarak belirlenir.

Önerilen yöntemin performansının değerlendirilmesi için AUROC kullanılır. Tablo 4.5'te verilen sınıf kategorileri ikiden fazla sınıfa sahip olduğundan her sınıf için AUROC skorları şöyle hesaplanır. Sınıf kategorisindeki sınıflardan biri pozitif sınıf kalan sınıflar ise negatif sınıf olarak belirlenir ve AUROC skoru hesaplanır. Bu işlem her bir sınıf için tekrarlanır ve AUROC skorlarının ortalaması, son AUROC skorunu verir.

Tablo 4.7 ve Tablo 4.8'de, sırasıyla Aminer ve HEP-Th veri kümelerinde üç ayrı nitelik kümesinin belirlenen sınıflandırıcılarda ayrı ayrı kullanılarak elde edilen AUROC sonuçları bulunur. Tablolardaki kalın değerler her bir sınıflandırıcının üç nitelik kümesindeki (F1, F5 ve F8) en iyi tahmin sonucunu gösterir ve altı çizili değerler her veri kümesi için tüm sınıflandırıcılar ve üç nitelik kümesi arasından elde en iyi sonucu gösterir. C4 ve C5 sınıf kategorileri için Aminer veri kümelerindeki en iyi AUROC skorları sırasıyla 0,832 ve 0,794'tür. Benzer şekilde, HEP-Th veri kümelerinde elde edilen en iyi sonuçlar C4 ve C5 için sırasıyla 0,842 ve 0,808'dir. C4 sınıf kategorisine göre örneklenen veri kümelerinin

Tablo 4.6. Örnekendirilmiş veri kümelerinin özellikleri

Örnekendirilmiş veri kümeleri		Test verisindeki sınıf örneklerinin toplam sayısı	w	n	Eğitim verisinin yılları	Test verisinin yılları	Zaman çerçeveleri kesişiyor mu?	Sınıf kategorisi
Aminer	D1_C4_FW1	12.000	1	4	2005-2008	2009	Hayır	C4
	D2_C4_FW2	12.000	2	3	2002-2007	2008-2009		
	D3_C4_FW2_I	12.000	2	5	2002-2007	2008-2009	Evet	
	D1_C5_FW1	11.560	1	4	2005-2008	2009	Hayır	C5
	D2_C5_FW2	15.000	2	3	2002-2007	2008-2009		
	D3_C5_FW2_I	15.000	2	5	2002-2007	2008-2009	Evet	
Hep-TH	D1_C4_FW1	9276	1	4	1999-2002	2003	Hayır	C4
	D2_C4_FW2	12.000	2	3	1996-2001	2002-2003		
	D3_C4_FW2_I	12.000	2	5	1996-2001	2002-2003	Evet	
	D1_C5_FW1	7810	1	4	1999-2002	2003	Hayır	C5
	D2_C5_FW2	15.000	2	3	1996-2001	2002-2003		
	D3_C5_FW2_I	15.000	2	5	1996-2001	2002-2003	Evet	

sonuçlarının C5'ten daha iyi olması beklenen bir sonuçtur, çünkü C4 kategorisindeki sınıfların atıf sayısı aralıkları C5'e göre daha geniştir, ayrıca C5'teki sınıf sayısı C4'ten bir fazladır. Problem, bağlantıların ağırlıklarıyla birlikte tahminini kapsadığından sadece bağlantıların tahmin edilmesinden daha zordur; bu açıdan önerilen bağlantı tahmin yaklaşımının performansı oldukça tatmin edicidir. Her iki veri kümesindeki testlerde RF ile çoğunlukla en iyi performans elde edilir. Bunu LR sınıflandırıcısı takip eder. Farklı zaman dilimlerine ve farklı pencere büyüklüğüne sahip veri kümelerinden elde edilen sonuçlar, önerilen yöntemin etkili olduğunu gösterir.

Tüm veri kümelerinde F5 nitelik kümesi F1'den daha iyi sonuç verir. Aminer ve HEP-Th veri kümelerinde F1'e göre F5 ile elde edilen maksimum performans artışı sırasıyla % 13 ve % 18 olarak dikkat çeker. F1 özelliği ($Score(v_s, v_t)$), birincil düğüm çifti ($P(v_s, v_t)$) ve ikincil düğüm çiftleri ($L(v_s, y)$, $L(y, v_s)$, $L(v_t, y)$, $L(y, v_t)$) arasındaki farklı bağlantı türlerinde yapılan hesaplamalar sonucu elde edilen tek bir puandır. Bu beş düğüm çifti F5 özelliğinde ayrı ayrı kullanılır. F5 ile elde edilen performans artışı, eğitimci bağlantı tahmininde farklı türdeki bağlantıların ayrı ayrı temsil edilmesiyle bağlantı oluşum mekanizmasının daha doğru öğrenilebileceğini gösterir.

F8 nitelik kümesiyle F5'e göre daha iyi sonuçlar elde edilir. Bu durum mevcut içerik tabanlı bilgilerin veya diğer topolojik ölçütlerin önerilen bağlantı tahmini metriği ile birlikte kullanılarak daha iyi performans elde edilebileceğini göstermektedir. Bununla birlikte, F5 ile elde edilen sonuçların F8 sonuçlarına yakın olması önerilen bağlantı tahmini ölçütünün etkinliğini doğrular.

Tablo 4.7. Aminer AUROC sonuçları

Veriler/Sınıflandırıcılar	IBK			RF			LR			MLP		
	F1	F5	F8	F1	F5	F8	F1	F5	F8	F1	F5	F8
D1_C4_FW1	0,748	0,786	0,781	0,726	0,814	0,826	0,760	0,788	0,820	0,764	0,783	0,819
D2_C4_FW2	0,749	0,764	0,789	0,721	0,797	0,822	0,787	0,798	0,826	0,789	0,790	0,820
D3_C4_FW2_I	0,755	0,788	0,781	0,737	0,817	0,832	0,779	0,802	0,827	0,784	0,791	0,822
D1_C5_FW1	0,706	0,731	0,745	0,687	0,771	0,791	0,735	0,756	0,787	0,738	0,752	0,785
D2_C5_FW2	0,719	0,744	0,740	0,694	0,784	0,794	0,751	0,770	0,789	0,755	0,765	0,787
D3_C5_FW2_I	0,708	0,742	0,736	0,683	0,778	0,791	0,749	0,771	0,793	0,754	0,765	0,790

Tablo 4.8. HEP-Th AUROC sonuçları

Veriler/Sınıflandırıcılar	IBK			RF			LR			MLP		
	F1	F5	F8	F1	F5	F8	F1	F5	F8	F1	F5	F8
D1_C4_FW1	0,732	0,784	0,782	0,716	0,829	0,842	0,778	0,817	0,822	0,780	0,817	0,818
D2_C4_FW2	0,721	0,753	0,746	0,699	0,793	0,802	0,768	0,794	0,799	0,771	0,794	0,796
D3_C4_FW2_I	0,711	0,749	0,740	0,696	0,789	0,799	0,764	0,791	0,796	0,766	0,792	0,793
D1_C5_FW1	0,685	0,740	0,734	0,669	0,794	0,808	0,743	0,787	0,793	0,743	0,782	0,787
D2_C5_FW2	0,668	0,702	0,700	0,644	0,752	0,765	0,730	0,757	0,763	0,735	0,757	0,761
D3_C5_FW2_I	0,675	0,709	0,705	0,658	0,754	0,766	0,731	0,759	0,763	0,733	0,760	0,761

4.3.3.2. Deney 2: Yeni Bağlantı Tahmini

Bağlantı tahmini problemi için önerilen yöntemlerin çoğu ağda ilk kez görülecek bağlantıların tahmini için test edilmiştir. Deney 2’de, önerilen ölçüt ve geleneksel bağlantı tahmini ölçütleri içinden etkili bir ölçüt olan AA ölçütü, etkili performansı [109, 46] çalışmalarında görülebilir, yeni bağlantı tahmini için test edilir. Bu testte verilen bir $G_{t,t'}(V, E)$ ağı için $G_{t'+1,t'+w}(V, E)$ ’de aralarında ilk kez bir atıf ilişkisi yoluyla bağlantı oluşacak bilim insanları tahmin edilir. Doğru bir karşılaştırma için [104]’de önerilen AA’nın zamansal ve ağırlıklı versiyonu (AA_WT) kullanılır. Önerilen bağlantı tahmin ölçütü ve AA_WT benzer zamansal yaklaşıma sahiptir, yakın zamanlarda oluşturulan bağlantıların skorları daha önceden oluşan bağlantılardan daha fazla ağırlıklandırılmıştır. AA_WT’de zaman ve ağırlık bilgilerini hesaplamaya katan fonksiyonlar kullanılır. Zamansal fonksiyon şu şekilde tanımlanır:

$$tf(v_i, v_j) = \log[(t(v_i, v_j) - t_{ilk}) + k] \quad (4.18)$$

burada t_{ilk} , $G_{t,t'}(V, E)$ ’de bağlantıların ilk oluşma zamanıdır; $t(v_i, v_j)$, v_i düğümünden v_j düğümüne doğru oluşan bağlantının zamanı gösterir ve k ise bir sabittir. Ağırlıklı zamansal fonksiyon, denklem 4.19’da verilir.

$$wtf(v_i, v_j) = w(v_i, v_j) \times tf(v_i, v_j) + w(v_j, v_i) \times tf(v_j, v_i) \quad (4.19)$$

Ağırlıklı zamansal fonksiyon, v_i 'den v_j 'ye ve v_j 'den v_i 'ye doğru oluşan bağlantıların ağırlık ve zaman niteliklerinin bir fonksiyonudur. AA_WT ölçütünün son şekli denklem 4.20'de tanımlanır.

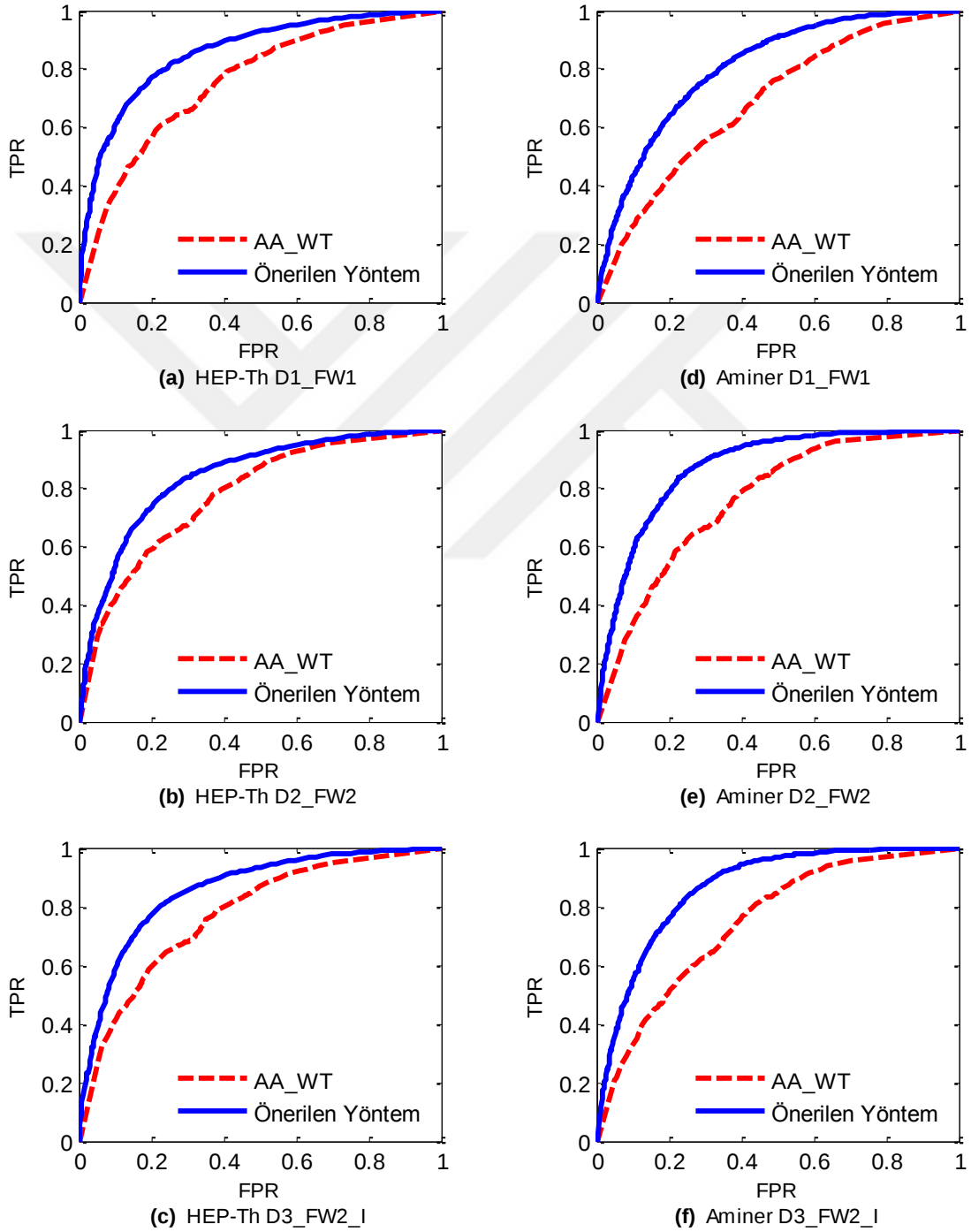
$$AA_WT(v_s, v_t) = \sum_{y \in I(v_s) \cap I(v_t)} \frac{wtf(v_s, y) + wtf(v_t, y)}{\log(1 + \sum_{c \in I(y)} wtf(y, c))} \quad (4.20)$$

$I(v_s)$, v_s 'nin komşu düğümlerinin kümesidir; y , v_s ve v_t 'nin ortak komşularıdır. Önerilen bağlantı tahmini ölçütünün F5 nitelik kümesi kullanılır. HEP-Th ve Aminer veri kümelerinin her biri için üç tane örneklendirilmiş veri kümesi oluşturulur. AA_WT komşuluk tabanlı bir ölçüttür, bu nedenle bu ölçüt eğitim verisinde 2 birim uzaklıktaki düğüm çiftleri için anlamlıdır. Doğru bir karşılaştırma yapmak için örnekler eğitim veri kümelerinde en az bir kez 2 birim uzaklıkta görülen düğüm çiftlerinden oluşturulur. Pozitif ve negatif olmak üzere iki sınıf belirlenir ve pozitif örnekler ağırlıklarına bakılmaksızın seçilir. Her bir örneklenmiş veri kümesi 5000 pozitif ve 5000 negatif örnekten oluşur. Örneklendirilmiş veri kümelerinin özellikleri Tablo 4.9'da listelenir. $wf(v_i, v_j)$ fonksiyonun k sabiti 3 olarak verilir. Aminer ve HEP-Th veri kümelerinde elde edilen AUROC sonuçları sırasıyla Tablo 4.10 ve 4.11'de verilir. Değerlendirme için ROC eğrileri de kullanılır. RF sınıflandırıcı ile en iyi AUROC sonuçları elde edildiğinden Şekil 4.3'te RF sınıflandırıcının sonuçlarının ROC eğrileri çizilir. Tablo 4.10 ve 4.11'deki kalın değerler her veri kümesi için en elde edilen en iyi sonucu gösterir. Önerilen bağlantı tahmini ölçütü tüm test veri kümelerinde AA_WT ölçütünden daha iyi performans sağlar. Önerilen zamansal olaylar sayesinde bağlantıların eğilimleri ağın gelişimi boyunca etkin bir şekilde belirlenir. AA_WT ölçütünde ise ağdaki eğilimler skora yansıtılamaz. AA_WT'de, v_s ve v_t 'nin yalnızca ortak komşuluk oluşturduğu durum dikkate

Tablo 4.9. Örneklendirilmiş veri kümelerinin özellikleri

Örneklendirilmiş veri kümeleri		Test verisindeki sınıf örneklerinin toplam sayısı	w	n	Eğitim verisinin yılları	Test verisinin yılları	Zaman çerçeveleri kesişiyor mu?
Aminer	D1_FW1	10.000	1	4	2005-2008	2009	Hayır
	D2_FW2	10.000	2	3	2002-2007	2008-2009	
	D3_FW2_I	10.000	2	5	2002-2007	2008-2009	Evet
HEP-Th	D1_FW1	10.000	1	4	1999-2002	2003	Hayır
	D2_FW2	10.000	2	3	1996-2001	2002-2003	
	D3_FW2_I	10.000	2	5	1996-2001	2002-2003	Evet

alınır. Önerilen ölçütte v_s ve v_t 'nin ortak komşuluk oluşturduğu durumun hesaplamaya katılmasının yanı sıra bu durumun öncesi ve sonrasında ağın zaman içerisinde evrimi boyunca bu düğüm çiftinin birbirlerine karşı artan veya azalan eğilimleri hesaplamaya katılır. Bu nedenlerden dolayı, önerilen ölçüt tüm testlerde AA_WT'yi geride bırakır.



Şekil 4.3. RF ile elde edilen sonuçların ROC eğrileri

Tablo 4.10. Aminer sonuçları

Veriler/Sınıflandırıcılar	IBK		RF		LR		MLP	
	AA	F5	AA	F5	AA	F5	AA	F5
D1_FW1	0,707	0,768	0,686	0,805	0,772	0,769	0,772	0,779
D2_FW2	0,784	0,855	0,761	0,871	0,834	0,836	0,834	0,838
D3_FW2_I	0,767	0,838	0,745	0,864	0,818	0,843	0,818	0,839

Tablo 4.11. HEP-Th sonuçları

Veriler/Sınıflandırıcılar	IBK		RF		LR		MLP	
	AA	F5	AA	F5	AA	F5	AA	F5
D1_FW1	0,774	0,821	0,754	0,856	0,824	0,828	0,824	0,842
D2_FW2	0,795	0,821	0,776	0,839	0,840	0,841	0,840	0,852
D3_FW2_I	0,793	0,826	0,771	0,856	0,842	0,844	0,842	0,853

4.4. Bölüm Değerlendirmesi ve Sonuçlar

Bu bölümde literatürde yeteri kadar üzerinde çalışılmayan bilim insanlarının atıf sayısı tahmini problemi için eğiticili bir öğrenme yöntemi önerilmiştir. Bilim insanlarının atıf sayısı tahmini atıf ağlarında bağlantı tahmini problemi olarak formüle edilmiştir. Atıf tahmini işi atıf ağlarında bağlantıların ağırlıklarıyla birlikte tahminini gerektirir. Bu açıdan bu bölümde önerilen yöntem yönlü, ağırlıklı ve zamansal atıf ağlarında bağlantıların ağırlıklarıyla birlikte tahmin edildiği ilk çalışmadır. Atıf ağlarının zaman içerisindeki evrimi ile artan/azalan eğilimlerin hesaplamaya katıldığı zamansal bir bağlantı tahmini ölçütü sunulmuştur. Önerilen ölçütte karmaşık ağlardaki yerel ve küresel topolojik yapılar kullanılmıştır.

Deney 1'de, önerilen zamansal bağlantı tahmini metriğinin iki versiyonu ve üç basit nitelik eğiticili bağlantı tahmini yönteminde kullanılmıştır. Deney 1, Aminer ve HEP-Th veri kümelerinden çıkarılan 12 örneklendirilmiş veri kümesi üzerinde gerçekleştirilmiştir. Doğru bir değerlendirme için her bir sınıfın örnek sayısı eşit tutulmuş ve negatif örnekler en fazla 5 birim uzaklıkta bulunan düğümler arasından seçilerek zorlu eğitim kümeleri oluşturulmuştur. Farklı sınıflandırıcılar tarafından elde edilen sonuçlar, bağlantıların ve bağlantıların ağırlıklarının tahminin zorluğu göz önünde bulundurulduğunda önerilen yöntemin performansının tatmin edici olduğunu gösterir. Elde edilen sonuçlar bilim dünyasındaki bibliyografik karmaşık ağlarda veri madenciliği çalışmalarının bilim

insanlarının etki tahmini veya bilimsel makalelerin etki tahmini için değerli bir çaba olduğunu da göstermiştir. F5 nitelik kümesi ile F1'e göre daha iyi sonuçların elde edilmesi, ağdaki farklı topolojik yapıların ölçümelerini ayrı ayrı temsil eden bir ölçüt oluşturma eğitici öğrenme yöntemlerinde daha etkili olduğunu göstermiştir. F8 nitelik kümesiyle çoğunlukla daha iyi sonuçlar elde edilmesi, uygun nitelik seçiminin bağlantı tahmini performansını iyileştirme potansiyeline sahip olduğunu göstermiştir. Daha iyi performans sonuçlarının elde edilebilmesi amacıyla bu bölümde önerilen zamansal bağlantı tahmini ölçütü ile bilim insanlarının makalelerindeki başlık, özet ve anahtar sözcükleri; bilim insanlarının görev yaptığı kuruluşlar; makalelerin yayınlandığı dergilerin etki faktörü; vb. gibi içerik temelli nitelikler ve diğer potansiyel topolojik niteliklerin eğitici bağlantı tahmini yöntemlerinde birlikte kullanılması sonraki bir çalışma olarak değerlendirilebilir.

Literatürde önerilen bağlantı tahmini ölçütlerinin çoğu, ağda görülecek yeni bağlantıların tahmininde test edilmiştir. Bu nedenle önerilen ölçüt, etkili bir geleneksel bağlantı tahmini ölçütü olan AA ölçütüyle Deney 2'de yeni bağlantı tahmini kapsamında karşılaştırılmıştır. Önerilen zamansal bağlantı tahmini ölçütü oluşturulan tüm örneklenmiş veri kümelerinde AA ölçütünden daha iyi performans göstermiştir. Deney 2'deki sonuçlar, zamansal karmaşık ağlardaki topolojik yapıların zenginliğinin ve bağlantıların zaman içerisindeki evriminin bağlantı tahmini için gelecek çalışmalara ışık tutabileceğini desteklemektedir.

5. KOMŞULUK TABANLI ÖLÇÜTLERİN YÖNLÜ, AĞIRLIKLILIK VE ZAMANSAL AĞLAR İÇİN GELİŞTİRİLMESİ

Karmaşık ağlarda bağlantı tahmini ilgi çeken problemlerden biridir. Literatürde bağlantı tahmini ile ilgili yapılan çalışmaların çok sayıda olması bu ilginin bir göstergesidir. Son zamanlarda yapılan çalışmalarda heterojen, zamansal veya yönlü ağ modelleriyle gerçek karmaşık ağlarda bağlantı tahmini için daha etkili yöntemler önerilmeye çalışılmıştır. Literatürdeki bağlantı tahmini ölçütlerinin çoğu, bağlantıların yön bilgisinin bağlantı oluşumundaki rolünü dikkate almamaktadır. Bu bölümde literatürde bağlantı tahmini için önerilen komşuluk tabanlı bağlantı tahmini ölçütlerinin etkinliğinin artırılması amacıyla bu ölçütleri yönlü ağlar için genişleten bir yöntem önerilir. Ayrıca daha etkili bağlantı tahmini yapılması için bağlantıların ağırlık ve zaman bilgisi de önerilen yöntemde hesaplama katılır. Önerilen yöntem Deney 1'de yönlü, ağırlıklı ve zamansal ağlarda literatürdeki dokuz tane komşuluk tabanlı bağlantı tahmini ölçütüyle eğitici öğrenme algoritmaları kullanılarak karşılaştırılır. Önerilen yaklaşımın performansı ağırlık ve zaman bilgilerini dikkate almayan yönlü ağlar için önerilmiş bir başka bağlantı tahmini ölçütü olan TC ölçütüyle Deney 2'de, test edilir. Deneysel sonuçlar önerilen yaklaşımın bağlantı tahmininin doğruluğunu önemli ölçüde artırdığını gösterir. Bu performans artışının daha fazla ağ bilgisi ve hesaplama zamanı gerektirmeden ağların topolojik niteliklerinin etkili bir şekilde kullanılmasıyla sağlanması önerilen yöntemin diğer bir katkı unsurudur.

Literatürdeki bağlantı tahmini yöntemlerinin çoğu, bağlantıların yönsüz olduğu ağlar için önerilmiştir. Örneğin komşuluk tabanlı bağlantı tahmini ölçütlerinde bağlantıların yön bilgileri hesaplamlarda ayırt edici bir unsur olarak kullanılmaz. Yol tabanlı ölçütlerde yön bilgisi kullanılabilmesine rağmen başlangıç düğüm noktasından hedef düğüm noktasına ulaşılamadığı yerel topolojik yapılar yol tabanlı ölçütlerde etkili bir şekilde hesaplama dahil edilemez. Literatürde az sayıdaki çalışmada etkili bağlantı tahmini için bağlantıların yönü dikkate alınmıştır. Örneğin, Schall [111] tarafından yönlü ağ motifleri kullanılarak eğitici bir bağlantı tahmini yöntemi sunulmuştur. Çalışmada ağ motifi tabanlı TC ölçütü önerilmiştir. TC, hesaplanacak düğüm çiftinin ağ motifiyle eşleşen kapalı üçlü motif sayısının, olası kapalı üçlü motif sayısı oranına dayanmaktadır. Lichtenwalter ve Chawla [112] tarafından bağlantı tahmini ve analizi için düğüm sıralama profili (Vertex Collocation Profile, VCP) olarak adlandırılan yeni bir kavram tanımlanmıştır. VCP, tahmin edilecek düğüm çiftlerini içeren tüm olası alt grafların vektörel gösterimidir, Vektörün her bir elemanı

olan alt graflar toplam n tane düğüm arasında r tane bağlantının olduğu yapılardır. n adet düğüm içerisine tahmin edilecek düğüm çifti de dahildir. VCP ile yönlü ve farklı türlere sahip bağlantılar gibi zengin ağ tabanlı yapılar ve ayrıca bağlantıların ağırlık ve zaman gibi nitelik bilgileri de sunabilir. Behfar vd. [113] bağlantı oluşum mekanizmalarını içe doğru bağlantılar (diğer düğümlerden söz konusu düğüme doğru gelen bağlantılar) ve dışa doğru bağlantılar (söz konusu düğümden diğer düğümlere doğru giden düğümler) üzerinden analiz etmişlerdir. Çalışma yönlü ağlarda içe doğru bağlantılar ve dışa doğru bağlantıların farklı derece dağılımlarına (degree distributions) sahip oldukları için bu yapıların farklı bağlantı oluşum mekanizmalarına sahip oldukları fikrine dayanmaktadır. Yapılan testlerde içe doğru bağlantı oluşumunda güç-kanunu dağılımının (power-law distribution), dışa doğru bağlantı oluşumunda ise ağır kuyruklu dağılımının (heavy-tailed degree distribution) izlendiği sonucuna varılmıştır. Shang vd. [114] yaptığı çalışmada bağlantı oluşum mekanizmasındaki tek yönlü ve çift yönlü bağlantıların farklı rollerini ortaya çıkarmak için yeni bir yönlü bağlantı tahmini yöntemi önerilmiştir. İki yönlü bir bağlantıyla birbirine bağlı düğüm çiftlerinin ortak komşularıyla bağlantı oluşturma ihtimalinin tek yönlü bir bağlantıyla birbirine bağlı düğüm çiftlerinden daha fazla olduğu sonucuna varılmıştır. Aghabozorgi ve Khayyambashi [115] tarafından bağlantıların farklı yönlerde dizilimlerini bulunduran üçlü ağ motiflerine dayalı yeni bir bağlantı tahmini ölçütü önerilmiştir. Önerilen ölçüt, sosyal ağlarda bağlantı tahmini için sınıflandırma modellerinde kullanılmıştır. Zhao vd. [115] tarafından potansiyel bağlantıların derecelendirilmesi için ağın topolojik yapıları ve bağlantıların ortak değişkenleri (covariates) kullanılarak yeni bir bağlantı tahmini yöntemi geliştirilmiştir. Yöntem hem yönlü hem de yönsüz ağlar için önerilmiştir, uç noktaları benzer olan düğüm çiftlerinin yönlü ağlarda bağlantı kurma olasılıklarının daha yüksek olduğu varsayımına dayanmaktadır. Ding ve Li [116] dünya ticaret ağlarındaki ticari hareketlerin tahmini için yönlü ağırlıklı ağların topolojik yapısını yeniden yapılandıran olasılık temelli bir yöntem önermişlerdir. Wang vd. [32], Zhou vd. [50] çalışmasını yönlü ağlarda bağlantı tahmini için geliştirmişlerdir. Ağın topolojik yapılarından mümkün olduğunca faydalanmak ve ağdaki bilgi kaybı sorununu azaltmak için orijinal ağa bir ekstra zemin düğümü eklenmiştir. Ağ eklenen bu zemin düğümünün ağdaki diğer tüm düğümlerle çift yönlü bağlantısı oluşturulmuştur. Önerilen yönlü bağlantı tahmini ölçütü, tahmin edilecek düğüm çiftleri arasındaki iki ve üç uzunluğundaki yolların sayısına ve eklenen zemin düğümü üzerinden gidilen üç uzunluğundaki yolların sayısına dayalıdır. Guo vd. [117] yönlü ağlarda bağlantıların yönünü tahmin etmek için düğümlerin derecelendirilmesine dayalı bir yöntem

önermişlerdir. Önerilen yöntem bağlantıların daha çok alt sıradaki düğümlerden daha üst sıradaki düğümlere doğru kurulduğu fikri üzerine geliştirilmiştir. Düğümlerin derecelendirilmesi için yerel ve küresel yapıları birleştiren özyinelemeli bir algoritma önerilmiştir.

Bu bölümde, yönlü ağlarda daha etkili bir bağlantı tahmini yapmak için komşuluk tabanlı ölçütler yönlü bağlantılardan oluşan ağ motifleri kullanılarak geliştirilir. Önerdiğimiz yöntemde bağlantıların ağırlık ve zaman nitelikleri de hesaplama katılır. [46, 64, 84, 118] çalışmalarında ağırlıklı/zamansal bağlantı tahmini ölçütleriyle daha etkili sonuçlar elde edildiği gösterilmiştir. Bu nedenle hem geleneksel komşuluk tabanlı ölçütler hem de önerdiğimiz motif tabanlı ölçüt [64] çalışmasındaki yöntemle ağırlıklı olarak kullanılmıştır. Ayrıca bu bölümde önerilen fonksiyonla ölçütler zamansal olarak uyarlanır. Önerilen yöntem ve geleneksel komşuluk tabanlı ölçütler eğitici bağlantı tahmini algoritmalarında karşılaştırılır. Elde edilen deneysel sonuçlar önerilen yöntemin etkinliğini ve uygulanabilirliğini gösterir.

5.1. Klasik Komşuluk Tabanlı Ölçütlerin Ağırlıklı ve Zamansal Olarak Genişletilmesi

Bu bölümde, Bölüm 2’de verilen komşuluk tabanlı ölçütlerin yönlü, ağırlıklı ve zamansal ağlar için uyarlaması verilir. Daha sonra bu ölçütlerle yönlü ağlarda daha etkili bağlantı tahmini yapılabilmesi için bu ölçütlerin yönlü bağlantılardan oluşan ağ motifleri kullanılarak genişletildiği yöntem verilir.

Bölüm 2’de verilen komşuluk tabanlı ölçütlerin hepsi ortak komşular bazında hesaplanmasına karşın bu ölçütler arasında bazı karakteristik farklılıklar vardır. Örneğin AA ve RA gibi ölçütlerde ağda seyrek ilişkilere sahip ve bu seyrek ilişkilerin her iki düğüm tarafından paylaşıldığı düğüm çiftleri daha yüksek şekilde derecelendirilir. HP ölçütünde hup rolünde olan düğümlere komşu olan düğüm çiftlerine daha yüksek skor değeri verilir. Bu farklılıklar, farklı karakteristik özelliklere sahip sosyal ağlarda her zaman belirli bir ölçütün en iyi bağlantı tahmini sonucunu veremeyebileceğini gösterir. Bu gerçek [109]’da sunulan çalışmada da gösterilmiştir. Bu açıdan bakıldığında tüm komşuluk tabanlı bağlantı tahmini ölçütlerin bağlantı tahmini performanslarını artırmak amacıyla bu bölümde önerdiğimiz genel yöntemin, bağlantı tahmini problemi için önemli bir katkı unsuru sağlayacağını gösterir.

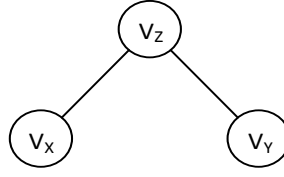
Komşuluk tabanlı bağlantı tahmini ölçütlerin ağırlıklı olarak genişletilmesi için De Sá ve Prudêncio'nun yaklaşımı [64] kullanılır. Denklem 5.1'de verilen fonksiyonla verilen v_x ve v_z düğüm çifti arasındaki bağlantının zamansal değeri hesaplanır.

$$tf(v_x, v_z) = \log[(t(v_x, v_z) - t_{ilk}) + k] \quad (5.1)$$

Bu denklemde verilen k parametresi bir sabittir, $t(v_x, v_z)$ v_x düğümünde v_z düğümüne doğru oluşan bağlantının ne zaman oluştuğu bilgisini verir. $tf(v_x, v_z)$ zamansal fonksiyonun bağlantı tahmini ölçütlerde kullanılmasıyla yakın zamanda oluşan bağlantılar daha fazla ağırlıklandırılır.

Tablo 5.1. Komşuluk tabanlı ölçütlerin yönlü ağırlıklı ve zamansal formları

Ölçütlerin denklemleri	
$AA_WT(v_x, v_y) =$	$\sum_{v_z \in I(v_x) \cap I(v_y)} \frac{wtf(v_x, v_z) + wtf(v_y, v_z)}{\log(1 + \sum_{v_c \in I(v_z)} wtf(v_z, v_c))}$
$CN_WT(v_x, v_y) =$	$\sum_{v_z \in I(v_x) \cap I(v_y)} wtf(v_x, v_z) + wtf(v_y, v_z)$
$HD_WT(v_x, v_y) =$	$\sum_{v_z \in I(v_x) \cap I(v_y)} \frac{wtf(v_x, v_z) + wtf(v_y, v_z)}{\max(\sum_{v_a \in I(v_x)} wtf(v_x, v_a), \sum_{v_b \in I(v_y)} wtf(v_y, v_b))}$
$HP_WT(v_x, v_y) =$	$\sum_{v_z \in I(v_x) \cap I(v_y)} \frac{wtf(v_x, v_z) + wtf(v_y, v_z)}{\min(\sum_{v_a \in I(v_x)} wtf(v_x, v_a), \sum_{v_b \in I(v_y)} wtf(v_y, v_b))}$
$JC_WT(v_x, v_y) =$	$\sum_{v_z \in I(v_x) \cap I(v_y)} \frac{wtf(v_x, v_z) + wtf(v_y, v_z)}{\sum_{v_a \in I(v_x)} wtf(v_x, v_a) + \sum_{v_b \in I(v_y)} wtf(v_y, v_b)}$
$LHN_WT(v_x, v_y) =$	$\sum_{v_z \in I(v_x) \cap I(v_y)} \frac{wtf(v_x, v_z) + wtf(v_y, v_z)}{\sum_{v_a \in I(v_x)} wtf(v_x, v_a) \times \sum_{v_b \in I(v_y)} wtf(v_y, v_b)}$
$RA_WT(v_x, v_y) =$	$\sum_{v_z \in I(v_x) \cap I(v_y)} \frac{wtf(v_x, v_z) + wtf(v_y, v_z)}{\sum_{v_c \in I(v_z)} wtf(v_z, v_c)}$
$SA_WT(v_x, v_y) =$	$\sum_{v_z \in I(v_x) \cap I(v_y)} \frac{wtf(v_x, v_z) + wtf(v_y, v_z)}{\sqrt{\sum_{v_a \in I(v_x)} wtf(v_x, v_a) \times \sum_{v_b \in I(v_y)} wtf(v_y, v_b)}}$
$SO_WT(v_x, v_y) =$	$\sum_{z \in I(x) \cap I(y)} \frac{2 \times (wtf(v_x, v_z) + wtf(v_y, v_z))}{\sum_{a \in I(x)} wtf(x, a) + \sum_{b \in I(y)} wtf(y, b)}$



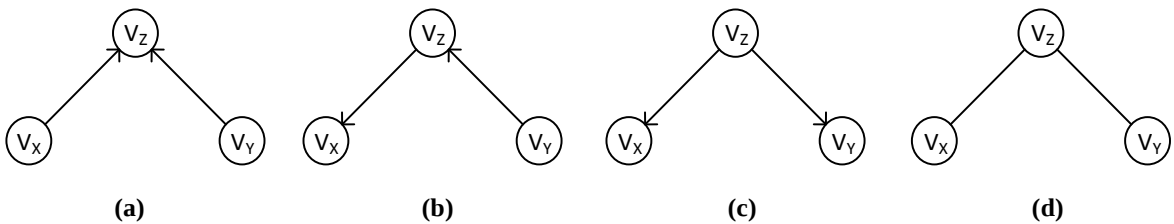
Şekil 5.1. Açık üçlü ağ motifi

Ölçütlerin zamansal ve ağırlık olarak düğüm çiftleri arasındaki yönlü bağlantıların hesaplanmasını sağlayacak fonksiyonun son hali denklem 5.2’de verilir. $wtf(v_x, v_z)$ fonksiyonu ile komşuluk tabanlı bağlantı tahmini ölçütler Tablo 5.1’de verildiği gibi ağırlıklı, zamansal ve yönlü ağlar için genişletilir.

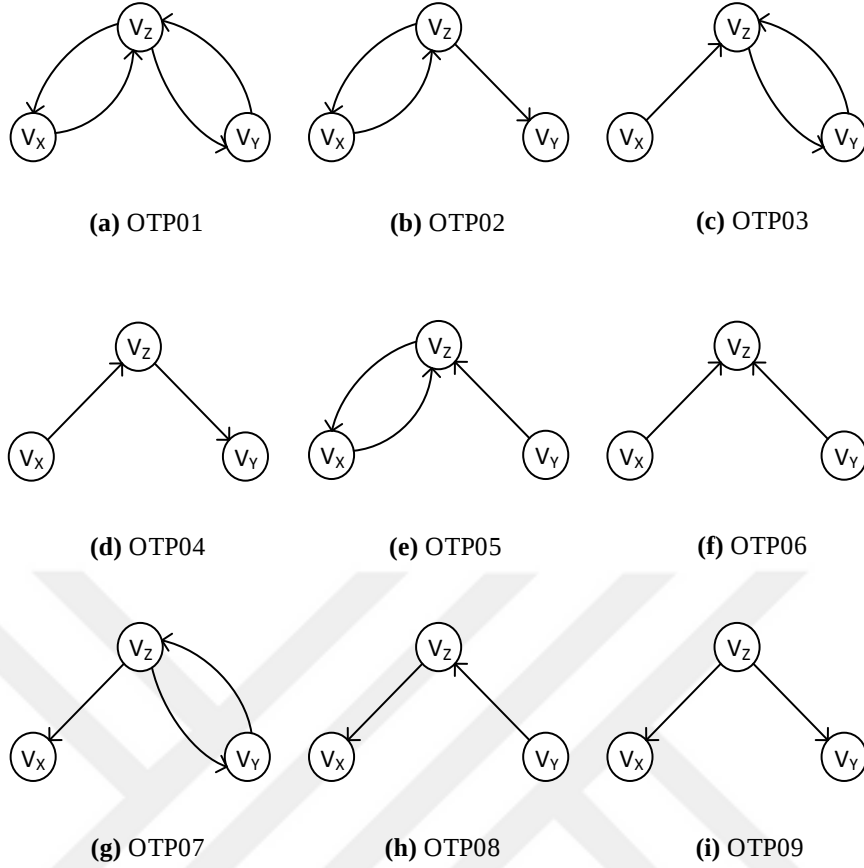
$$wtf(v_x, v_z) = w(v_x, v_z) \times tf(v_x, v_z) + w(v_z, v_x) \times tf(v_z, v_x) \quad (5.2)$$

Önerdiğimiz yöntem yönlü ağlarda test edileceğinden doğru bir karşılaştırma yapılması amacıyla klasik komşuluk tabanlı ölçütler de yönlü ağlar için hesaplanır. Bu hesaplama dışarı ve içeri yönlü bağlantıların tümünün işleme katılmasıyla yapılır.

Önerilen yöntemin kavramsal yapısı için açık üçlü ağ motifi (open triad pattern, OTP) olarak adlandırılan yerel bir topolojik yapı tanımlanır. Şekil 5.1’de gösterildiği gibi bir $OTP(v_x, v_y, v_z)$ yapısı, v_x , v_y ve v_z olmak üzere üç düğümden ve (v_x, v_z) ve (v_y, v_z) düğüm çiftleri arasındaki bağlantılardan oluşur. Bir OTP’de (v_x, v_y) düğüm çifti arasında bağlantı yoktur, v_z düğümü v_x ve v_y düğümünün ortak komşusudur. OTP, v_x ve v_y düğümleri arasında bir bağlantı olup olmayacağının araştırılması için kullanılır. Tablo 5.1’de listelenen tüm geleneksel komşuluk tabanlı bağlantı tahmini ölçütlerinde OTP yapısı kullanılır. Denklemlerde $z \in I(x) \cap I(y)$ olarak tanımlanan her bir düğüm v_x ve v_y düğümleriyle birlikte bir $OTP(v_x, v_y, v_z)$ oluşturur. Yönsüz ağlarda Şekil 5.1’de gösterilen tek bir OTP tipi v_x ve v_y düğüm çiftleri için kullanılabilir. Ancak yönlü ağlarda farklı türde OTP’ler



Şekil 5.2. Yönlü (a-c) ve yönsüz (d) bağlantılardan oluşan bir kaç OTP örneği



Şekil 5.3. OTP türleri

mümkündür. Yönlü OTP yapıları, yönlü ağlar için başka çalışmalarda da [111, 119] kullanılmıştır. Bir sonraki bölümde, düğüm çiftleri için bağlantıların farklı yönlerde olduğu OTP'ler ve komşuluk tabanlı ölçütlerin yönlü bağlantılar için genişletildiği yöntem verilir.

5.2. Komşuluk Tabanlı Ölçütlerin Yönlü Ağlar İçin Geliştirilmesi

Yönlü ağlarda belirli bir düğüm çifti (v_x, v_y) için farklı yönlü bağlantılardan oluşan OTP'ler, Tablo 5.1'de verilen komşuluk tabanlı ölçütler tarafından ayırt edici şekilde hesaplamaya katılmazlar. Örneğin komşuluk tabanlı ölçütler Şekil 5.2 (d)'de gösterilen sadece bir OTP'yi kullanabilirler, Şekil 5.2 (a-c) 'de aynı düğümlerin farklı yönlerdeki bağlantılarla oluşturduğu farklı OTP'ler için aynı puanları hesaplar. Ancak, Şekil 5.2 (a-c)'de farklı eğilimler bulunur. v_z düğümü, Şekil 5.2 (a) 'da v_x ve v_y düğümlerinin ilgisini çekerken Şekil 5.2 (c)' de durum tam tersidir. Şekil 5.2 (b)'de, v_x düğümü v_z düğümünün doğrudan ilgisini çekerken v_y düğümünün dolaylı şekilde ilgisini çeker. Bu çeşitli durumlar

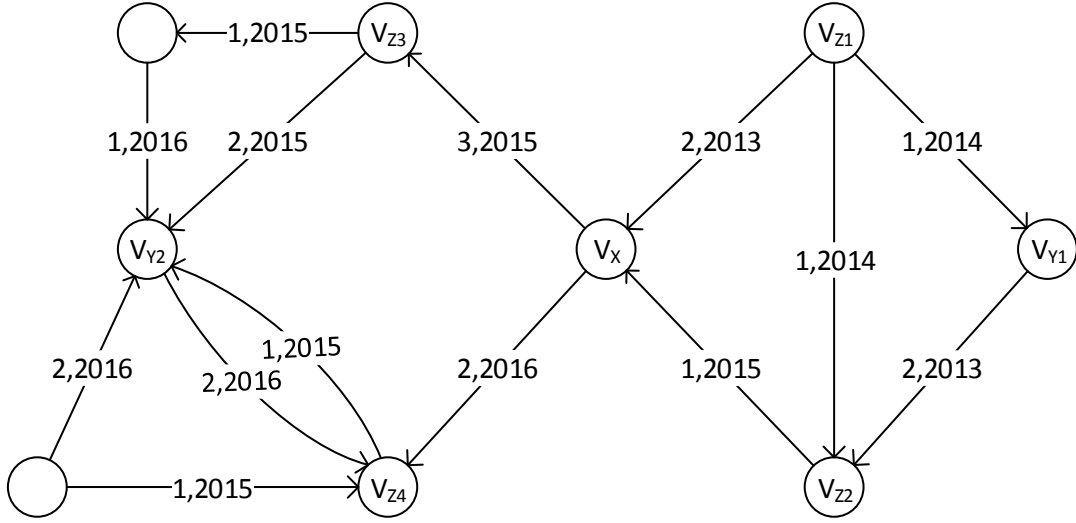
bağlantı oluşumunda farklı etkilere sahiptir. Klasik komşuluk tabanlı ölçütlerde yönlü bağlantılar sadece dışarı yönlü ya da sadece içeri yönlü bağlantılar şeklinde hesaplamaya katılabilir. Ancak bu durumda hem içeri hem dışarı yönlü bağlantıların oluşturduğu yerel yapılar kullanılamayacağından bir dezavantaj oluşturur. Diğer bir yaklaşım ise komşuluk tabanlı ölçütlerin yönlü ağlarda hem içeri hem de dışarı yönlü bağlantıların birlikte hesaplamaya katılmasıdır. Bu yaklaşım önceki bölümde Tablo 5.1’de verilen komşuluk tabanlı ölçütlerin ağırlıklı yönlü ve zamansal olarak genişletildiği yaklaşımdır. Ancak yine de klasik komşuluk tabanlı ölçütler yönlü OTP’lerde bağlantı oluşum mekanizmalarını etkin bir şekilde tespit edemez.

Klasik komşuluk tabanlı ölçütlerle yönlü ağlarda daha etkili bağlantı tahmini yapılabilmesi amacıyla bu ölçütler önerdiğimiz yönlü ağ motiflerine dayalı bir yaklaşımla genişletilir. Yönlü ağlardaki olası tüm OTP tipleri Şekil 5.3’te gösterilir. Yönlü komşuluk tabanlı ölçüt için önerdiğimiz EXT_NLPM’nin genel tanımını şöyledir:

$$\text{EXT_NLPM}(v_x, v_y) = \{skor_{OTP1}, skor_{OTP2}, \dots, skor_{OTP9}\} \quad (5.3)$$

EXT_NLPM komşuluk tabanlı ölçütlere karşılık gelen bir OTP skorlar kümesidir. EXT_NLPM hesaplanırken her bir $v_z \in \Gamma(v_x) \cap \Gamma(v_y)$ düğümü için öncelikle v_x , v_y ve v_z düğümlerinin oluşturduğu OTP türü tespit edilir. Daha sonra Tablo 5.1’deki tüm komşuluk tabanlı ölçütlerin skorları v_x , v_y ve v_z düğümleri için hesaplanır, tür bilgisi ve skor değeri kaydedilir. v_x ve v_y düğümleriyle aynı OTP türüne sahip v_z düğümlerinin skor değerleri toplanır. OTP türlerinin bulunması için eğitim veri kümesindeki tüm bağlantılar kullanılır. Eğitim veri kümesi zaman çerçeveleriyle bölünmüş ağlara dönüştürülerek bir zamansal hesaplama yapılabilirdi. Ancak bu durumda bölünmüş farklı ağlarda bağlantıları olan OTP türleri tespit edilemeyecekti. Bu nedenle eğitim veri kümesinin başlangıç ve bitiş zaman aralığındaki tüm bağlantılar OTP türünün tespiti için kullanılır. Zamansal hesaplama için ise önerilen $tf(v_x, v_z)$ fonksiyonunda yararlanılır.

Şekil 5.4’te EXT_NLPM’nin hesaplanmasının gösterimi için örnek bir ağ verilir. Bu ağdaki (v_x, v_{y1}) ve (v_x, v_{y2}) düğüm çiftleri için klasik komşuluk tabanlı ve EXT_NLPM ölçüt değerleri Tablo 5.2’de verilir. Örnek ağdaki yönlü bağlantılar üzerindeki değerler sırasıyla ağırlık ve zaman nitelik değerlerini gösterir.



Şekil 5.4. EXT_NLPM hesaplamasının gösterimi için örnek bir ağ

Tablo 5.2. Şekil 5.4'te verilen örnek ağdaki (v_x, v_{y1}) ve (v_x, v_{y2}) düğüm çiftleri için klasik komşuluk tabanlı ve önerilen yönlü komşuluk tabanlı ölçüt değerleri

Komşuluk tabanlı ölçütler	(v_x, v_{y1})		(v_x, v_{y2})	
	Standart değerleri	EXT_NLPM değerleri	Standart değerleri	EXT_NLPM değerleri
AA_WT	4,08	{0, 0, 0, 0, 0, 0, 0, 2,08, 2}	7	{0, 0, 0, 3,6, 3,4, 0, 0, 0, 0}
CN_WT	7,38	{0, 0, 0, 0, 0, 0, 0, 3,8, 3,58}	16,81	{0, 0, 0, 8,77, 8,04, 0, 0, 0, 0}
HD_WT	0,6	{0, 0, 0, 0, 0, 0, 0, 0,31, 0,29}	1,21	{0, 0, 0, 0,63, 0,58, 0, 0, 0, 0}
HP_WT	2,06	{0, 0, 0, 0, 0, 0, 0, 1,06, 1}	1,36	{0, 0, 0, 0,71, 0,65, 0, 0, 0, 0}
JC_WT	0,46	{0, 0, 0, 0, 0, 0, 0, 0,24, 0,22}	0,63	{0, 0, 0, 0,33, 0,3, 0, 0, 0, 0}
LHN_WT	0,16	{0, 0, 0, 0, 0, 0, 0, 0,08, 0,08}	0,09	{0, 0, 0, 0,05, 0,04, 0, 0, 0, 0}
RA_WT	1,45	{0, 0, 0, 0, 0, 0, 0, 0,73, 0,72}	1,67	{0, 0, 0, 0,84, 0,83, 0, 0, 0, 0}
SA_WT	1,11	{0, 0, 0, 0, 0, 0, 0, 0,57, 0,54}	1,29	{0, 0, 0, 0,67, 0,62, 0, 0, 0, 0}
SO_WT	0,93	{0, 0, 0, 0, 0, 0, 0, 0,48, 0,45}	1,28	{0, 0, 0, 0,67, 0,61, 0, 0, 0, 0}

Hesaplama $tf(v_x, v_z)$ fonksiyonun k parametresi 3 olarak belirlenir, ağdaki bağlantılar 2013 ve 2016 yılları arasında oluşur. (v_x, v_{y1}) ve (v_x, v_{y2}) düğüm çiftleri farklı OTP türlerine sahipken klasik komşuluk tabanlı ölçütler bu düğüm çiftleri için sadece bir skor değeri üretir. Bu skor değeri de farklı OTP'leri yansıtmaz. Ancak önerilen EXT_NLPM ölçütü (v_x, v_{y1}) düğüm çifti için OTP8 ve OTP9 skor değerlerini, (v_x, v_{y2}) düğüm çifti için OTP4 ve OTP5 skor değerlerini üretir. Böylece karmaşık ağlardaki topolojik yapılardan bağlantı tahmini için daha çok kazanım elde edilir. Ağırlık ve zaman bilgileri komşuluk tabanlı ölçütler ve EXT_NLPM'lerde bağlantıların nitelikleri olarak kullanılır. Ağırlık bilgileri kullanılarak

düğümler arasındaki ilişkilerin kuvveti dikkate alınır. Ağırlık bilgisinin önemi Şekil 5.4'te gösterilen örnek ağda görülebilir. Örneğin, (v_x, v_{y1}) ve (v_x, v_{y2}) düğüm çiftlerinin ağırlıklı CN skorları sırasıyla 6 ve 9'dur, bunların ağırlıksız skoru ise her iki düğüm için 2'dir. Ağırlıklı CN puanları, v_{y2} ile v_x arasındaki ilişkinin v_{y1} 'e göre daha kuvvetli olduğunu gösterir. Yakın zamanda oluşan bağlantıların sonraki bağlantı oluşumlarında daha belirleyici olduğu varsayımı genel olarak kabul edilen bir varsayımdır. Bu durum Şekil 5.4'te verilen örnek ağda da tasvir edilir. Örneğin (v_x, v_{y1}) ve (v_x, v_{y2}) düğüm çiftlerinin zamansal olmayan CN puanı her ikisi için 2 iken, bu düğüm çiftlerinin zamansal CN puanları sırasıyla 5,193 ve 6,804'tür. Zamansal CN puanları $tf(v_x, v_z)$ fonksiyonu kullanılarak hesaplanır. Elde edilen bu farklı skorlar v_{y2} 'nin v_{y1} 'e göre v_x ile arasında daha yakın zamanlı ilişkilerin olduğunu gösterir.

5.3. Eğitici Bağlantı Tahmini

Verilen bir $G_{t,t'}(V, E)$ ağı için $[t'', t''']$ ($t'' > t'$) zamanında iki düğüm arasında yeni bir bağlantının oluşup oluşmayacağı problemi ikili bir sınıflandırma problemi olarak formüle edilir. Sınıflandırma işlemi WEKA [110] kullanılarak gerçekleştirilir. Klasik komşuluk tabanlı ölçütlerin ve bunların yönlü olarak genişletildiği EXT_NLPM ölçütlerin performansları sınıflandırıcılarda ayrı ayrı test edilir. Bağlantı tahmininin doğruluğunun artırılması amacıyla EXT_NLPM ölçütleri ve belirlenen birkaç nitelik sınıflandırıcılarda birlikte kullanılır.

5.4. Deneyler

5.4.1. Deney Konfigürasyonları

Sınıflandırma işlemi için K-En Yakın Komşu, Rastgele Orman, Rastgele Ağaç, Çok Katmanlı Algılayıcılar ve C4.5 sınıflandırıcılar kullanılır. Bunların WEKA'daki karşılıkları sırasıyla IBk, RandomForest (RF), RandomTree (RT), MultilayerPerception (MLP) ve J48'dir. RF, MLP ve J48 için varsayılan parametreler ayarlanır. IBk için KNN parametresi 5, RT için kat sayısı (numfolds) parametresi 3 olarak ayarlanır. Kullanılan sınıflandırıcıların tahmin hatası 10 kat çapraz doğrulama ile değerlendirilir.

5.4.2. Test Veri Kümeleri ve Ağ Modeli

Önerilen yöntemin performansı DBLP ve HEP-Th veri kümelerinde test edilir. DBLP ve HEP-Th veri kümelerinden iki atıf ağı oluşturulur. Bölüm 4.3.2’de kullanılan ağ modeli bu bölümde de kullanılır. Bu ağ modelinde düğümler bilim insanlarını, yönlü bağlantılar bilim insanları arasındaki atıfları gösterir. Yönlü bağlantılar ağırlık ve yıl nitelik bilgilerini içerir. Ağ modelinin örnek bir gösterimi Bölüm 4.3.2’te bulunur. Oluşturulan atıf ağlarının özellikleri Tablo 5.3’te verilir.

Tablo 5.3. HEP-Th ve DBLP atıf ağlarının özellikleri

Veri kümesi	Bağlantılar	Düğümmler	Yıllar
HEP-Th	735.710	8380	1992-2003
DBLP	61.660	5695	1970-2001

Karmaşık ağlar çok seyrek yapıda olduğundan dolayı pozitif örneklerin sayısı negatif örneklerin sayısına göre oldukça azdır. Bu durum sınıf dengesizliğine yol açar ve önerilen bağlantı tahmini yöntemini performansının doğru değerlendirilmesini etkiler. Bu nedenden dolayı örnekleme yöntemi kullanılır. Eğitim veri setlerinde her yıl için en az 2 kez atıf alan bilim insanları içinden pozitif ve negatif sınıf örnekleri rastgele ve eşit sayıda seçilir. Pozitif sınıf örnekleri test veri kümesinde aralarında ilk defa bağlantı oluşan düğüm ikililerinden, negatif sınıf örnekleri ise test veri kümesinde aralarında bağlantı olmayan düğüm ikililerinden seçilir. Pozitif ve negatif örnekler eğitim veri ağına aralarında en fazla 2 birim uzaklıktaki düğüm çiftlerinden seçilir. Negatif örnekler uzaklığa bakılmaksızın rastgele seçilseydi gerçek ağların seyrek yapısından dolayı düğüm çiftlerinin çoğu büyük olasılıkla aralarında uzak mesafe olan düğüm çiftlerinde oluşacaktı. Pozitif örneklere benzeyen negatif örnekleri birbirinden ayırt etmek sınıflandırma işleminin en kritik noktasıdır. Aksi takdirde negatif olduğu kuvvetle muhtemel bir verinin negatif olduğunun tahmini, önerilen yöntemin etkin olduğunun anlaşılmasında yeterli değildir. Her bir sınıf için örnek sayısının eşit sayıda tutulmasıyla, pozitif ve negatif örneklerin 2 birim uzaklıkta bulunan ve her yıl en az 2 atıf alan düğüm çiftlerinden seçilmesiyle önerilen yöntemin performansının doğru bir şekilde değerlendirilebileceği eğitim verileri oluşturulur.

5.4.3. Deney Sonuçları

Bu bölümde yapılan çalışmanın esas amacı yönlü ağlarda bağlantıların yön bilgisinin bağlantı tahminin doğrulunu artırıp artırmadığının araştırılmasıdır. Bu nedenle Deney 1'de komşuluk tabanlı ölçütler önerilen EXT_NLPM ölçütüyle karşılaştırılır. Deney1'de komşuluk tabanlı ölçütler ve bu ölçütlerin yönlü ağlar için genişletildiği EXT_NLPM ölçütleri ağırlıklı, zamansal ve yönlü ağlar için hesaplanır. Deney 2'de, EXT_NLPM'nin bağlantı tahmini performansı ve yönlü bir bağlantı tahmini ölçütü olan TC ölçütü [10] ile karşılaştırılır. TC, ağırlıksız ve zamansal olmayan ağlar için önerildiğinden doğru bir karşılaştırma olması amacıyla EXT_NLPM'nin ağırlıksız ve zamansal olmayan formu Deney 2'de kullanılır.

5.4.3.1. Deney 1: Komşuluk Tabanlı Ölçütlerin EXT_NLPM Ölçütleri ile Yönlü, Ağırlıklı ve Zamansal Ağlarda Karşılaştırılması

HEP-Th ve DBLP veri kümelerinden sırasıyla 6000 ve 2000 örnek veri pozitif ve negatif örneklerin sayısı eşit olacak şekilde seçilir. Örneklendirilmiş veri kümelerinin ayrıntıları Tablo 5.4'te sunulur. DBLP ve HEP-Th veri kümelerinin AUROC değerleri Tablo 5.5'te verilir. DBLP veri kümesindeki deney sonuçlarında her bir komşuluk tabanlı ölçütün performansının EXT_NLPM ile iyileştirildiği görülür. SO_WT ölçütünün AUROC sonucu EXT_SO_WT ile %28 artışla geliştirilir. Ayrıca EXT_SO_WT ölçütüyle sınıflandırıcılar içinde SO_WT'ye kıyasla %20'nin üzerinde bir değerle en yüksek ortalama artış sağlanır. En iyi AUROC değeri (0,805) RF sınıflandırıcı tarafından EXT_AA_WT ile elde edilir. Sınıflandırıcılarda EXT_NLPM ölçütüyle elde edilen sonuçlara bakıldığında RF daha sonra IBk ile çoğunlukla en iyi performansın sağlandığı görülür. DBLP veri kümesinde her bir sınıflandırıcı için komşuluk tabanlı ölçütlere karşı EXT_NLPM ölçütüyle elde edilen en yüksek performans sonuçlarının ve en yüksek performans artışının sağlandığı sonuçların ROC eğrileri Şekil 5.5 ve Şekil 5.6'da gösterilir. Tablo 5.5'te verilen HEP-Th veri kümesinin AUROC sonuçlarında da görüldüğü gibi EXT_NLPM ölçütü ile genişletilen komşuluk tabanlı ölçütlerin hepsi klasik komşuluk tabanlı ölçütleri geride bırakmıştır. SO_WT ölçütünün AUROC sonucu EXT_SO_WT ile %65 gibi dikkat çeken bir oranda artırılır. Ayrıca EXT_LHN_WT ile sınıflandırıcılar içinde LHN_WT'ye karşı %46 en yüksek ortalama artış değeri elde edilir. En iyi AUROC değeri (0,854) MLP sınıflandırıcı tarafından

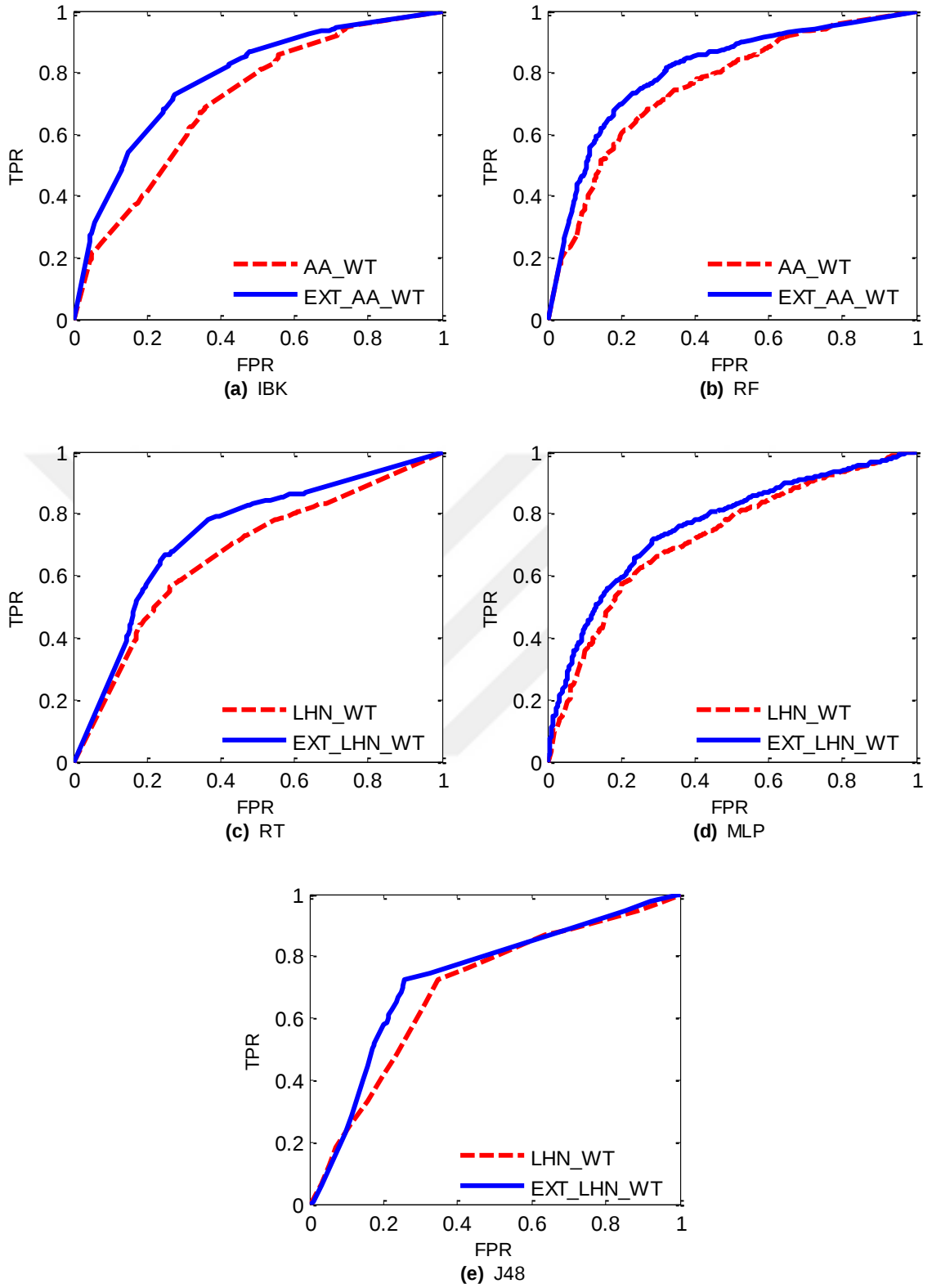
Tablo 5.4. Örneklendirilmiş HEP-Th ve DBLP veri kümelerinin özellikleri

Veri kümeleri	Eğitim veri kümesindeki bağlantı sayısı	Eğitim veri kümesindeki düğüm sayısı	Eğitim veri kümelerinin yılları	Test verisindeki pozitif ve negatif örneklerin sayısı	Test verilerinin yılları
HEP-TH	250.000	4700	1999-2002	6000	2003
DBLP	10.000	2400	1994-1997	2000	1998

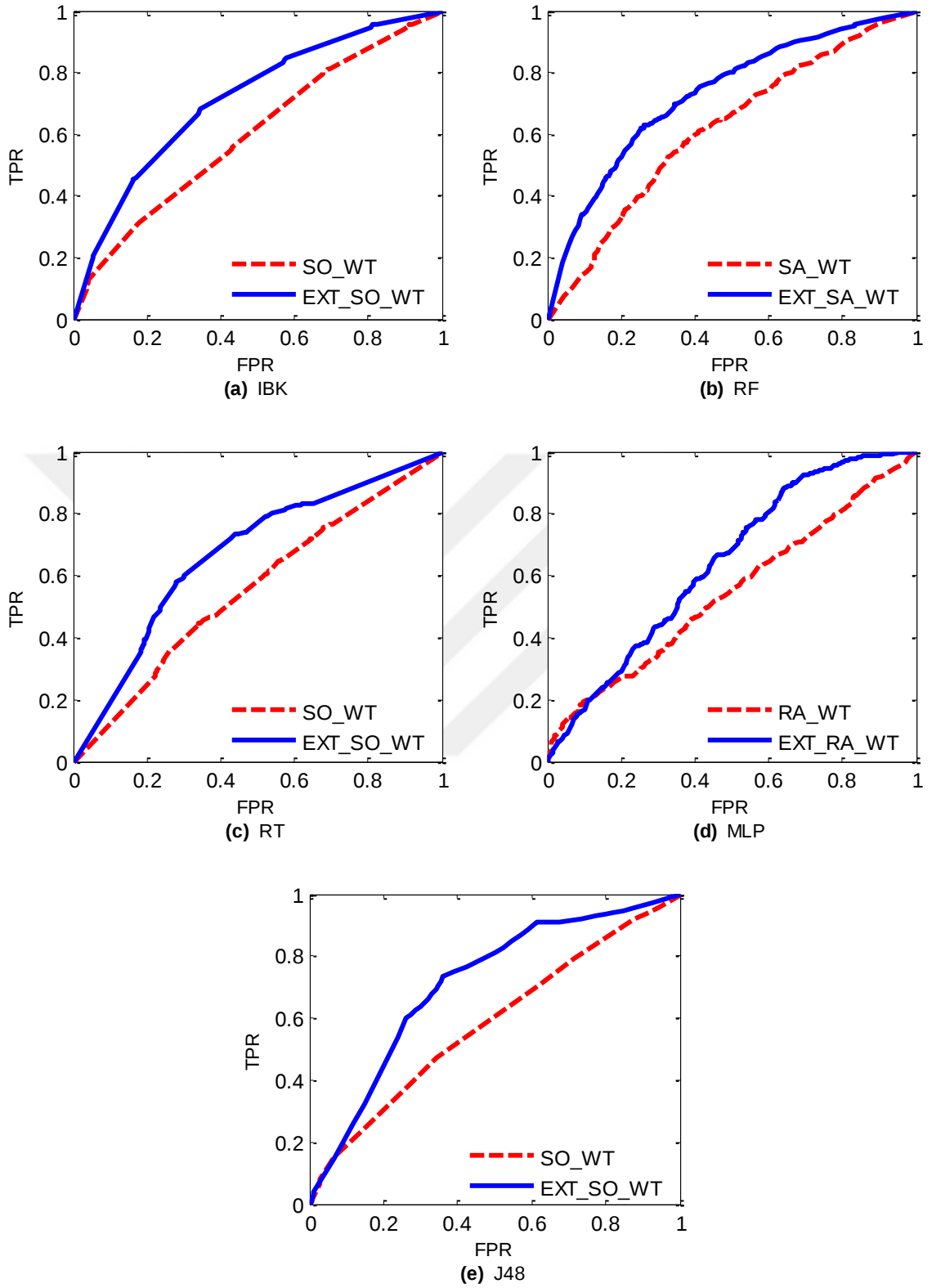
Tablo 5.5. DBLP ve HEP-Th veri kümelerinin AUROC sonuçları

	IBK		RF		RT		MLP		J48	
	DBLP	HEP-	DBLP	HEP-	DBLP	HEP-	DBLP	HEP-	DBLP	HEP-
AA_WT	0,709	0,806	0,756	0,782	0,693	0,747	0,661	0,839	0,610	0,777
EXT_AA_WT	0,782	0,829	0,805	0,842	0,722	0,772	0,733	0,854	0,725	0,814
CN_WT	0,725	0,805	0,728	0,772	0,696	0,750	0,682	0,850	0,662	0,779
EXT_CN_WT	0,770	0,828	0,776	0,844	0,729	0,775	0,751	0,851	0,730	0,812
HD_WT	0,640	0,714	0,664	0,687	0,609	0,659	0,601	0,772	0,560	0,707
EXT_HD_WT	0,733	0,806	0,769	0,823	0,705	0,754	0,711	0,824	0,702	0,816
HP_WT	0,647	0,752	0,670	0,727	0,610	0,709	0,641	0,809	0,625	0,752
EXT_HP_WT	0,751	0,814	0,759	0,838	0,698	0,774	0,730	0,829	0,715	0,824
JC_WT	0,639	0,728	0,638	0,696	0,574	0,689	0,692	0,790	0,640	0,726
EXT_JC_WT	0,721	0,807	0,737	0,828	0,676	0,777	0,713	0,836	0,708	0,814
LHN_WT	0,701	0,581	0,697	0,560	0,671	0,557	0,761	0,499	0,688	0,511
EXT_LHN_WT	0,762	0,789	0,771	0,817	0,738	0,765	0,783	0,780	0,733	0,807
RA_WT	0,692	0,655	0,720	0,635	0,633	0,617	0,544	0,517	0,572	0,597
EXT_RA_WT	0,754	0,817	0,786	0,838	0,716	0,778	0,686	0,854	0,708	0,830
SA_WT	0,596	0,783	0,614	0,753	0,575	0,735	0,632	0,831	0,579	0,766
EXT_SA_WT	0,709	0,807	0,729	0,830	0,672	0,768	0,726	0,850	0,716	0,818
SO_WT	0,600	0,729	0,623	0,697	0,557	0,695	0,614	0,790	0,563	0,726
EXT_SO_WT	0,715	0,804	0,719	0,821	0,671	0,771	0,717	0,833	0,722	0,813

EXT_AA_WT ve EXT_RA_WT ile elde edilir. İki veri kümesinden elde edilen deneysel sonuçlar AA ölçütünün en iyi bağlantı tahmini ölçütlerinden biri olduğunu gösterir, bu ölçütün başarılı performansı çalışmalarında da gösterilmiştir [109, 46]. Önerilen yöntem sayesinde AA ölçütünün başarılı performansı geliştirilir. Sınıflandırıcıların performansına bakıldığında çoğunlukla RF daha sonra MLP ile en iyi performans elde edilir. HEP-Th veri kümesinde EXT_NLPM ile elde edilen başarılı performanslar, Şekil 5.7 ve Şekil 5.8'deki ROC eğrilerinde de görülür. Her sınıflandırıcıda EXT_NLPM ölçütüyle elde edilen en iyi performansın ve buna karşılık gelen klasik komşuluk tabanlı ölçüt performansının ROC eğrileri Şekil 5.7'de çizilir. Şekil 5.8'de her sınıflandırıcıda en yüksek performans artışının



Şekil 5.5. DBLP’de en iyi EXT_NLPM ölçütü sonucunun ve buna karşılık gelen komşuluk tabanlı ölçütü sonucunun ROC eğrileri

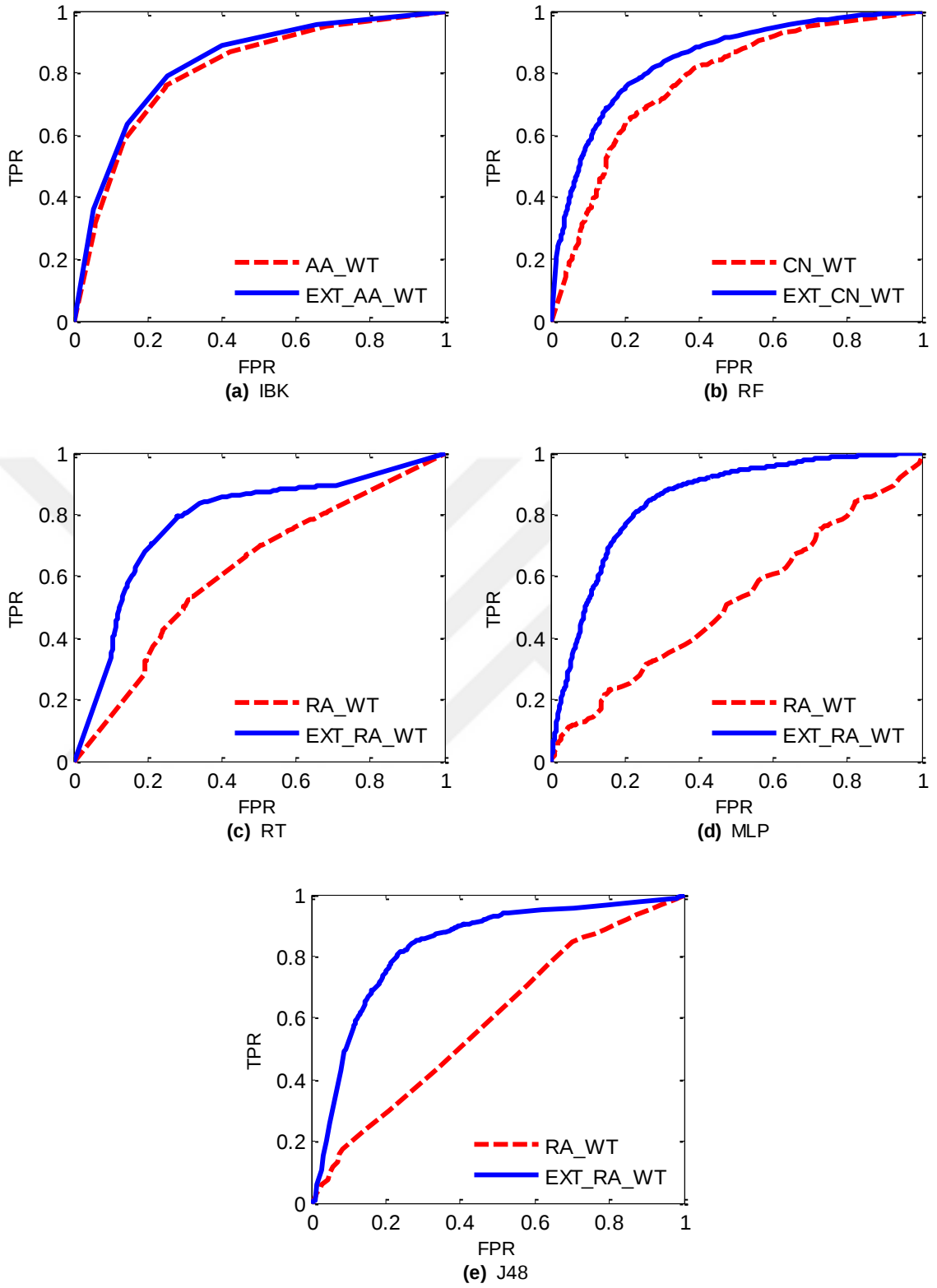


Şekil 5.6. DBLP’de komşuluk tabanlı ölçütlere göre en yüksek performans artışının sağlandığı sonuçların ROC eğrileri

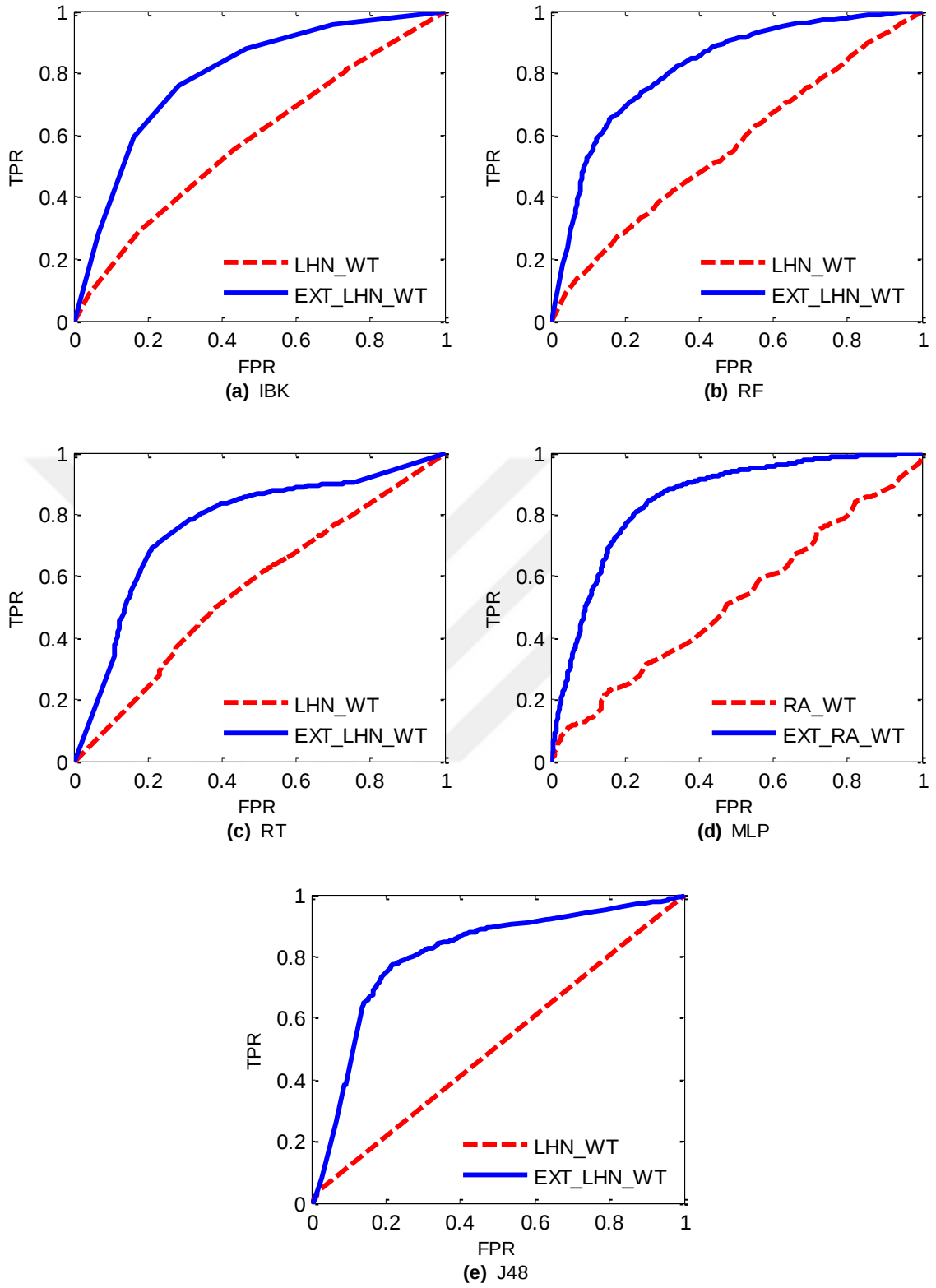
elde edildiği EXT_NLPM ölçütünün ve buna karşılık gelen komşuluk tabanlı ölçütün ROC eğrileri gösterilir. DBLP ve HEP-Th veri kümeleri kullanılarak yapılan deneyler önerilen yaklaşımın etkinliğini doğrular. Bu başarılı performans şöyle açıklanabilir. Önerilen EXT_NLPM ölçütünde verilen düğüm çiftinin eşleştiği yönlü ağ motifine göre komşuluk tabanlı ölçüt hesaplanır, belirlenen dokuz tane ağ motifi için bu hesaplama sonucunda dokuz tane skor değeri elde edilir. Klasik komşuluk tabanlı ölçütlerde ise verilen düğüm çifti için bağlantıların yön bilgilerinin ayırt edilmediği bir skor değeri elde edilir. Bu yüzden önerilen yaklaşımla komşuluk tabanlı ölçütlerin yönlü ağlarda bağlantı tahmini doğruluğu önemli ölçüde artırılır.

Komşuluk tabanlı ölçütler düşük hesaplama karmaşıklığına sahiptir, çünkü bu ölçütler yerel ölçütlerdir, karmaşık ağlardaki düğüm çiftlerinin ortak komşuları üzerinden hesaplanırlar. Komşuluk tabanlı ölçütlerin zaman karmaşıklığı [35] 'te verilmiştir ve şöyle tanımlanmıştır. Düğümlerin ortalama komşu sayısının n olduğu bir karmaşık ağda bir düğümün tüm komşularına ulaşmanın zaman karmaşıklığı $O(n)$ 'dir. v_x ve v_y düğüm çiftinin komşularının kesişim ya da birleşim kümesinin elde edilmesi $O(n^2)$ zamanda olur. CN, HD, HP, LHN, SA, SO ölçütlerinin zaman karmaşıklığı $O(n^2)$ 'dir, çünkü bu ölçütler v_x ve v_y düğümlerin komşularının kesişimi üzerinden hesaplanır. JC ölçütünde v_x ve v_y düğümlerin komşularının kesişim ve birleşim kümesine ihtiyaç duyulur, bu yüzden bu ölçütün hesaplanması $O(2n^2)$ zaman alır. AA ve RA'nın zaman karmaşıklığı $O(2n^2)$ 'dir, çünkü bu ölçütlerde her biri $O(n^2)$ zamanını alan iki işlemin hesaplanması gerekir. Birincisi v_x ve v_y düğümünün her bir ortak komşusunun komşularının bulunması, ikincisi ise v_x ve v_y düğümün ortak komşularının bulunmasıdır. EXT_NLPM yönteminin zaman karmaşıklığı komşuluk tabanlı ölçütlerle aynıdır, çünkü EXT_NLPM'de komşuluk tabanlı ölçütlerde zaman alan işlemlerin dışında ayrıca zaman alan bir işlem yoktur. EXT_NLPM yönteminde komşuluk tabanlı ölçütlerden farklı olarak (v_x, v_y) düğüm çiftinin ortak komşularıyla oluşturduğu ağ motifi türünün belirlenmesi işlemi vardır. Bu işlem v_x ve v_y düğümlerin ortak komşularının bulunduğu esnada düğümler arasındaki bağlantıların yönleri üzerinde yapılan karşılaştırmalarla yapılır. Dolayısıyla EXT_NLPM yöntemi de düşük hesaplama karmaşıklığına sahiptir.

Karmaşık ağların bazı temel topolojik niteliklerinin bağlantı tahmini performansı üzerindeki etkisinin incelenmesi için EXT_NLPM ölçütü ve bazı temel topolojik nitelikler sınıflandırıcılarda birlikte kullanılır. EXT_NLPM+ olarak adlandırılan bu nitelik kümesi



Şekil 5.7. HEP-Th’de en iyi EXT_NLPM ölçütü sonucunun ve buna karşılık gelen komşuluk tabanlı ölçütü sonucunun ROC eğrileri

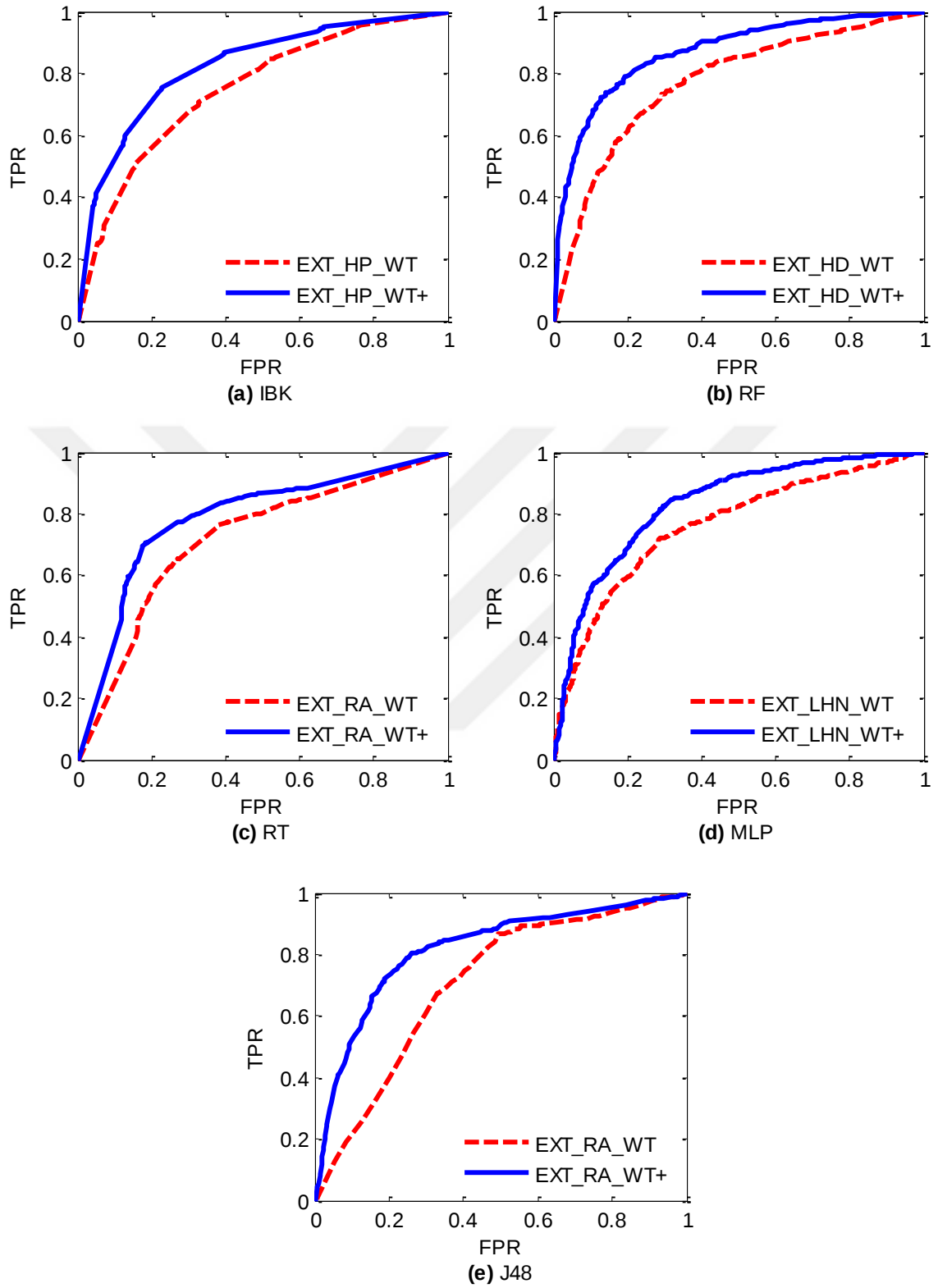


Şekil 5.8. HEP-Th’de komşuluk tabanlı ölçütlere göre en yüksek performans artışının sağlandığı sonuçların ROC eğrileri

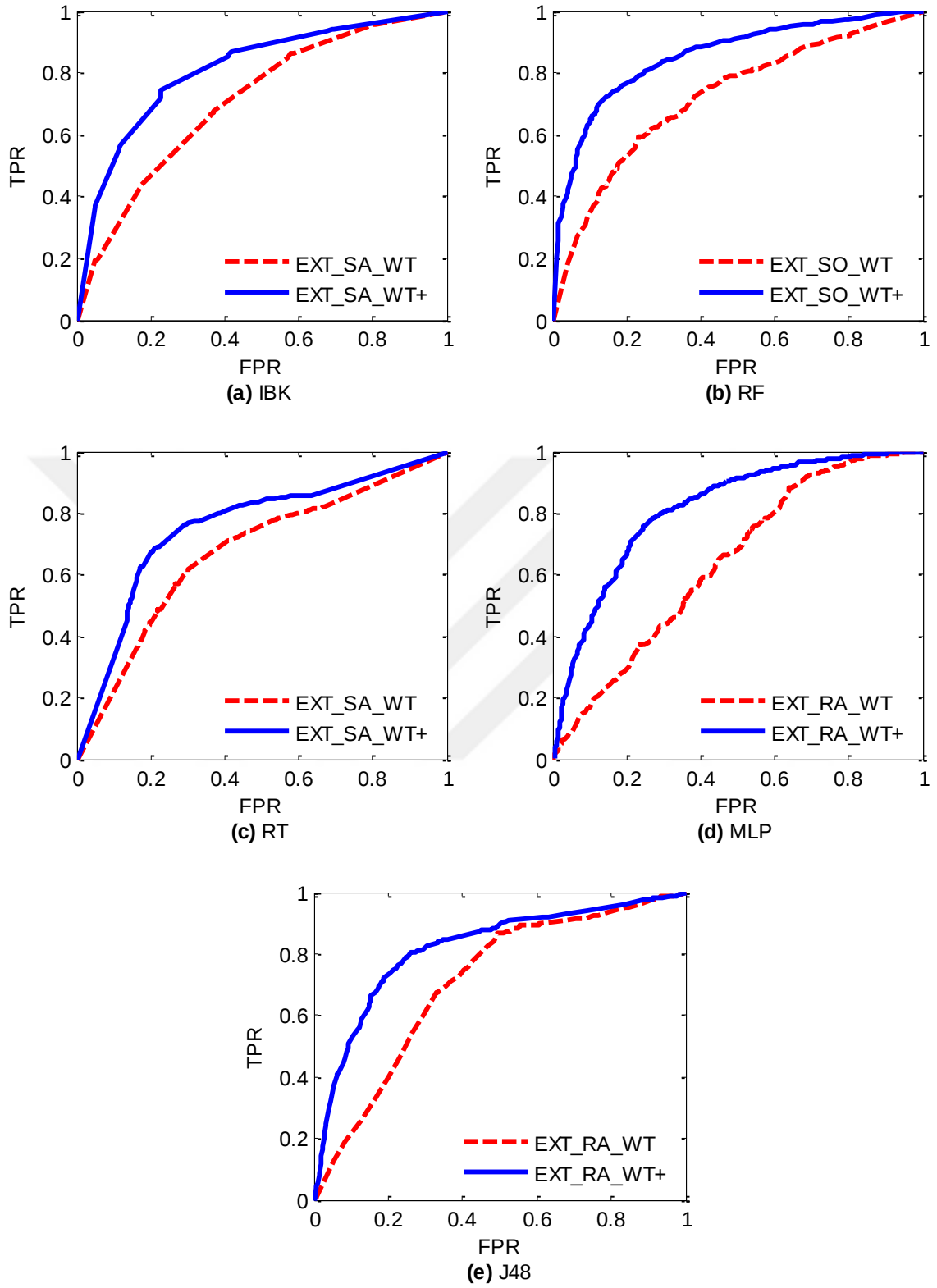
9 elemanlı EXT_NLPM nitelik kümesi, v_y 'nin toplam atıf sayısı ve v_y 'nin atıflarının kaç farklı yazardan olduğu bilgisi olmak üzere 11 elemandan oluşur. DBLP ve HEP-Th veri kümelerinde EXT_NLPM ve EXT_NLPM+ nitelik kümelerinin karşılaştırma sonuçları Tablo 5.6'da verilir. DBLP veri kümesinde EXT_NLPM+'nin tüm sonuçları EXT_NLPM'ye kıyasla daha yüksek performans gösterir. Maksimum performans artışı %19 ile EXT_RA_WT+ 'den elde edilir. En iyi AUC değeri (0,869) RF sınıflandırıcı tarafından EXT_HD_WT+ ile elde edilir. RF sınıflandırıcısıyla tüm EXT_NLPM+ deneyleri içinde en iyi performans elde edilir. DBLP veri kümesinde her bir sınıflandırıcı için EXT_NLPM'ye karşı EXT_NLPM+ ile elde edilen en yüksek performans sonuçlarının ve en yüksek performans artışının sağlandığı sonuçların ROC eğrileri Şekil 5.9 ve Şekil 5.10'da gösterilir. HEP-Th veri kümesinde EXT_CN_WT+ nitelik kümesiyle RT sınıflandırıcıda elde edilen sonuç hariç tüm EXT_NLPM+ sonuçları EXT_NLPM'den daha iyi performans gösterir. MLP sınıflandırıcı tarafından EXT_LHN_WT+ ile %6 en iyi performans artışı sağlanır. Yine MLP sınıflandırıcı ve EXT_RA_WT+ nitelik kümesiyle 0,878 en iyi AUROC değerine ulaşılır. HEP-Th veri kümesinde EXT_NLPM+ nitelik kümesiyle iyileştirilmiş performans sonuçlarının ROC eğrileri Şekil 5.11 ve Şekil 5.12'de gösterilir.

Tablo 5.6. DBLP ve HEP-Th veri kümelerinin EXT_NLPM ve EXT_NLPM+ AUROC sonuçları

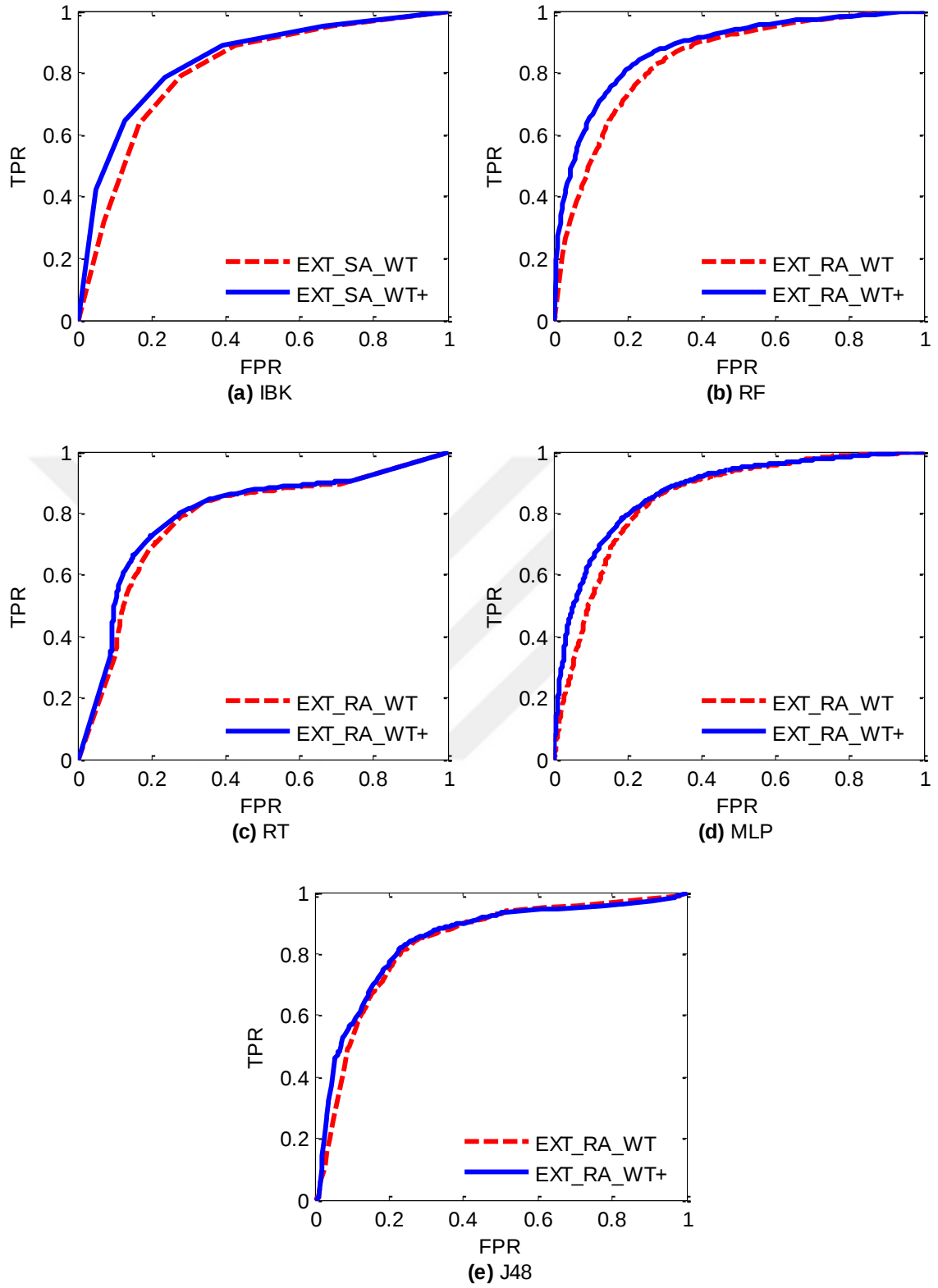
	IBK		RF		RT		MLP		J48	
	DBLP	HEP-Th	DBLP	HEP-Th	DBLP	HEP-Th	DBLP	HEP-Th	DBLP	HEP-Th
EXT_AA_WT	0,782	0,829	0,805	0,842	0,722	0,772	0,733	0,854	0,725	0,814
EXT_AA_WT+	0,814	0,837	0,858	0,872	0,782	0,776	0,828	0,874	0,804	0,826
EXT_CN_WT	0,770	0,828	0,776	0,844	0,729	0,775	0,751	0,851	0,730	0,812
EXT_CN_WT+	0,813	0,835	0,854	0,869	0,754	0,774	0,828	0,871	0,802	0,826
EXT_HD_WT	0,733	0,806	0,769	0,823	0,705	0,754	0,711	0,824	0,702	0,816
EXT_HD_WT+	0,803	0,828	0,869	0,868	0,753	0,774	0,824	0,865	0,804	0,826
EXT_HP_WT	0,751	0,814	0,759	0,838	0,698	0,774	0,730	0,829	0,715	0,824
EXT_HP_WT+	0,825	0,824	0,853	0,871	0,747	0,781	0,826	0,860	0,796	0,830
EXT_JC_WT	0,721	0,807	0,737	0,828	0,676	0,777	0,713	0,836	0,708	0,814
EXT_JC_WT+	0,805	0,837	0,853	0,872	0,756	0,785	0,830	0,872	0,799	0,832
EXT_LHN_WT	0,762	0,789	0,771	0,817	0,738	0,765	0,783	0,780	0,733	0,807
EXT_LHN_WT+	0,808	0,818	0,857	0,868	0,764	0,783	0,834	0,833	0,799	0,816
EXT_RA_WT	0,754	0,817	0,786	0,838	0,716	0,778	0,686	0,854	0,708	0,830
EXT_RA_WT+	0,805	0,837	0,867	0,877	0,783	0,793	0,819	0,878	0,815	0,843
EXT_SA_WT	0,709	0,807	0,729	0,830	0,672	0,768	0,726	0,850	0,716	0,818
EXT_SA_WT+	0,811	0,838	0,858	0,875	0,755	0,781	0,824	0,869	0,795	0,826
EXT_SO_WT	0,715	0,804	0,719	0,821	0,671	0,771	0,717	0,833	0,722	0,813
EXT_SO_WT+	0,811	0,834	0,854	0,870	0,744	0,788	0,831	0,864	0,801	0,830



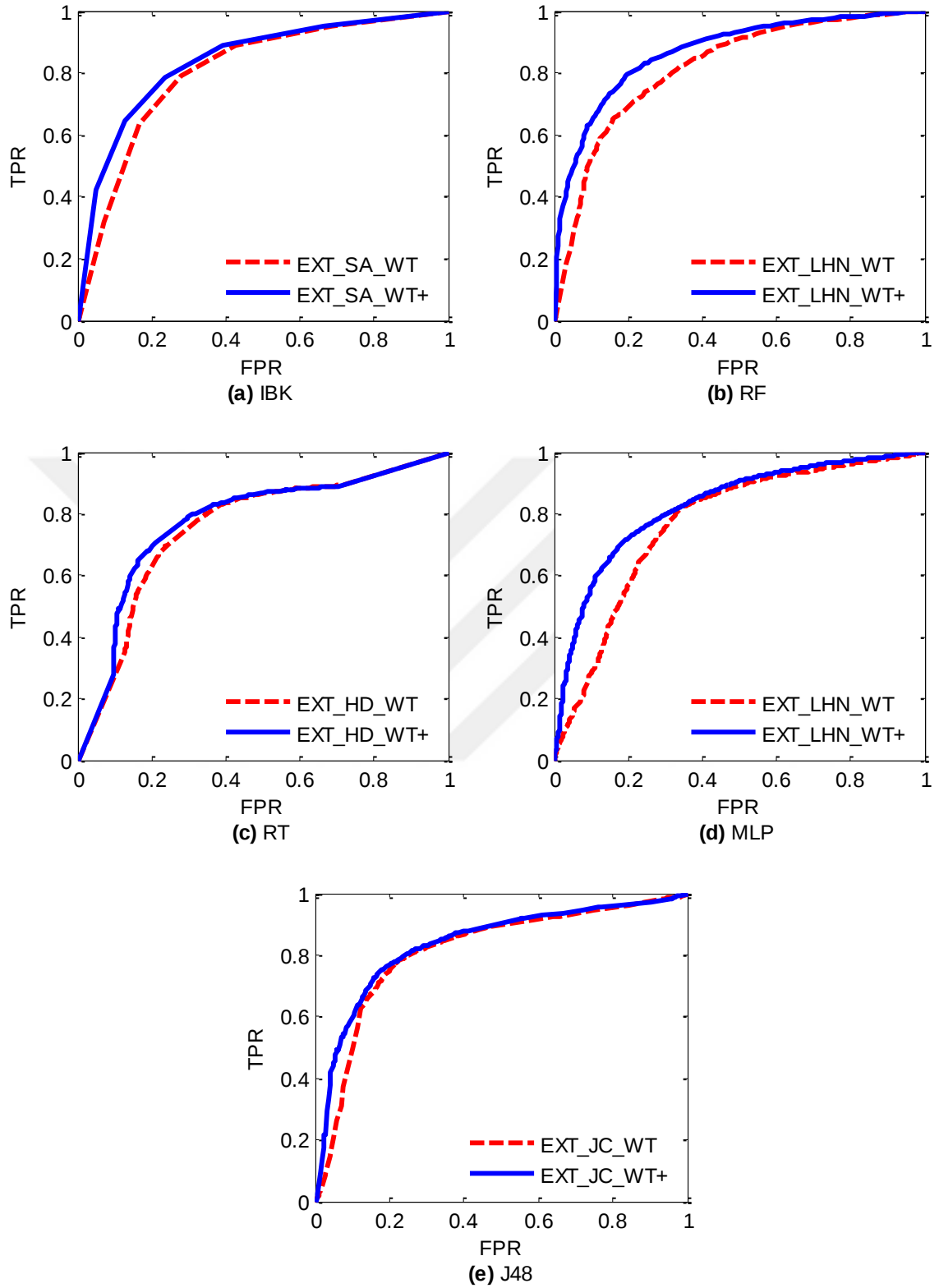
Şekil 5.9. DBLP’de en iyi EXT_NLPM+ sonucunun ve buna karşılık gelen EXT_NLPM sonucunun ROC eğrileri



Şekil 5.10. DBLP’de EXT_NLPM’ye karşı EXT_NLPM+ ile en yüksek performans artışının sağlandığı sonuçların ROC eğrileri



Şekil 5.11. HEP-Th'de en iyi EXT_NLPM+ sonucunun ve buna karşılık gelen EXT_NLPM sonucunun ROC eğrileri



Şekil 5.12. HEP-Th’de EXT_NLPM’ye karşı EXT_NLPM+ ile en yüksek performans artışının sağlandığı sonuçların ROC eğrileri

5.4.3.2. Deney 2: EXT_NLPM ile TC'nin Yönlü Ağlarda Karşılaştırılması

Bu bölümde, önerilen EXT_NLPM yöntemi ile yönlü ağlar için yakın zamanda önerilmiş TC ölçütü Deney 2'de karşılaştırılır. TC, ağırlıksız ve zamansal olmayan ağlar için önerildiğinden doğru bir performans karşılaştırması yapılması amacıyla EXT_NLPM ölçütleri de ağırlıksız ve zamansal olmayan şekliyle kullanılır.

TC ölçütü, bağlantıların yönlü olduğu üç düğümden oluşan ağ motiflerinin karmaşık ağlarda görülme sıklığına dayalı bir bağlantı tahmini ölçütüdür. Bölüm 5.2'de verilen OTP yapıları TC ölçütünde de kullanılmıştır. TC ölçütüyle hangi OTP'lerin kapalı OTP formu oluşturacağını tahmini yapılmaya çalışılmıştır. Bir OTP üç farklı şekilde kapalı forma dönüşebilir: birinci durum v_x düğümünden v_y düğüme doğru bir bağlantının oluşmasıdır; ikinci durum birinci durumun tam tersidir, v_y düğümünden v_x düğüme doğru bir bağlantının oluşmasıdır ve son durum hem v_x 'ten v_y 'ye doğru hem de v_y 'den v_x 'ye doğru bağlantının oluşmasıdır. Bu üç kapalı OTP türleri sırasıyla CTP1X, CTP2X ve CTP3X olarak isimlendirilir. Her bir kapalı üçlü ağ motifi (CTP) türünde dokuz OTP türünün karşılığı vardır, örneğin CTP11, ... CTP19. TC ölçütü verilen bir düğüm çiftinin aralarındaki ortak komşularıyla oluşturduğu OTP türüyle eşleşen CTP türünün ağdaki toplam sayısının, söz konusu OTP türünün ağdaki toplam sayısına oranıyla hesaplanır. v_x ve v_y düğümleri için TC ölçütü şöyle tanımlanmıştır:

$$TC(v_x, v_y) = \sum_{v_z \in I(v_x) \cap I(v_y)} w^p(v_x, v_z, v_y) * w(v_z) \quad (5.4)$$

$w^p(v_x, v_z, v_y)$ fonksiyonu üç düğümün oluşturduğu OTP ve CTP'lerin ağda görülme sıklığına dayalıdır ve şöyle tanımlanmıştır:

$$w^p(v_x, v_z, v_y) = \frac{F(T(v_x, v_z, v_y) + 10) + F(T(v_x, v_z, v_y) + 30)}{F(T(v_x, v_z, v_y))} \quad (5.5)$$

$T(v_x, v_z, v_y)$ fonksiyonu v_x , v_y ve v_z düğümlerin oluşturduğu OTP türünü verir ve çıktısı OTP0X formatındadır. $F(T(v_x, v_z, v_y) + 10)$ ve $F(T(v_x, v_z, v_y) + 30)$ fonksiyonları, sırasıyla OTP0X tipine karşılık gelen CTP1X ve CTP3X türlerinin eğitim verilerinde görülme sıklığını verir. TC ölçütünde $w(v_z)$ fonksiyonu ile AA ve RA ölçütlerinde olduğu

gibi ağda nadir görülen ilişkileri ortak paylaşan düğüm çiftlerinin skorları ağırlıklandırılmıştır. Bu fonksiyon şöyle tanımlanmıştır:

$$w(v_z) = \frac{1}{k_{vz}} \quad (5.6)$$

Pozitif ve negatif örneklerin sayısı eşit olacak şekilde HEP-Th ve DBLP veri kümeleri içinden 2000 veri örneklendirilir. Örnekleme Bölüm 5.4.2’de detayları verilen örnekleme ölçülerine göre yapılır. Örneklendirilmiş veri kümelerinin detayları Tablo 5.7’de verilir. DBLP ve HEP-Th veri kümelerinin AUROC sonuçları sırasıyla Tablo 5.8 ve Tablo 5.9’da

Tablo 5.7. Örneklendirilmiş veri kümelerinin özellikleri

Veri kümeleri	Eğitim veri kümesindeki bağlantı sayısı	Eğitim veri kümesindeki düğüm sayısı	Eğitim veri kümelerinin yılları	Test verisindeki pozitif ve negatif örneklerin sayısı	Test verisinin yılı
HEP-TH	443.614	7143	1999-2002	2000	2003
DBLP	21.591	3252	1994-1997	2000	1998

verilir. Her iki veri kümesinde elde edilen sonuçlarda EXT_NLPM ölçütlerinin TC ölçütünü büyük çoğunlukla geride bıraktığı görülür. EXT_NLPM ölçütlerinin daha üstün bağlantı tahmini performansı şöyle açıklanabilir. Aralarında bağlantı oluşup oluşmayacağı tahmin edilen düğüm çiftleri ortak komşularıyla karmaşık ağlarda çoğunlukla çeşitli OTP formlarında görülürler. TC ölçütünde farklı OTP türleriyle eşleşen CTP türlerine göre hesaplanır. Ancak bu ölçütte bağlantı oluşumunda hangi OTP türünün daha belirleyici olduğu etkin bir şekilde belirlenemez. Çünkü farklı OTP türlerinin skorları bir arada toplanır. Önerdiğimiz EXT_NLPM yönteminde ise her bir OTP türünün skor değerinin ayrı ayrı tutulduğu bir skor kümesi oluşturulur. Bu skor kümesi eğitici bağlantı tahmini modellerinde kullanarak bağlantı oluşumunda hangi OTP türünün daha belirleyici olduğu etkili bir şekilde öğrenilir.

DBLP veri kümesinde çoğunlukla EXT_HP ve EXT_RA ölçütleriyle en iyi bağlantı tahmini performansını sağlarken, HEP-Th veri kümesinde EXT_AA ölçütüyle çoğunlukla en iyi performans elde edilir. Deney 1 ve Deney 2’nin sonuçları, EXT_AA ve EXT_RA’nın en güçlü bağlantı tahmini ölçütleri olduğunu gösterir. Ancak bununla birlikte her zaman en iyi bağlantı tahmini performansını veren bir ölçüt sonuçlarda gözlenmez. Bu durum

önerdiğimiz yöntemi daha da önemli hale getirmektedir; çünkü önerilen yöntemde komşuluk tabanlı bağlantı tahmini ölçütlerin tümü bağlantıların yön bilgileri de dikkate alınarak genel bir yaklaşımla genişletilir ve bu sayede daha iyi performans elde edilir. Her iki veri kümesinde, RF sınıflandırıcı en iyi AUROC değerini verir. RF sınıflandırıcısında EXT_NLPM ve TC ile elde edilen bağlantı tahminini sonuçları, Şekil 5.13'de ROC eğrileriyle de gösterilir.

5.5. Bölüm Değerlendirmesi ve Sonuçlar

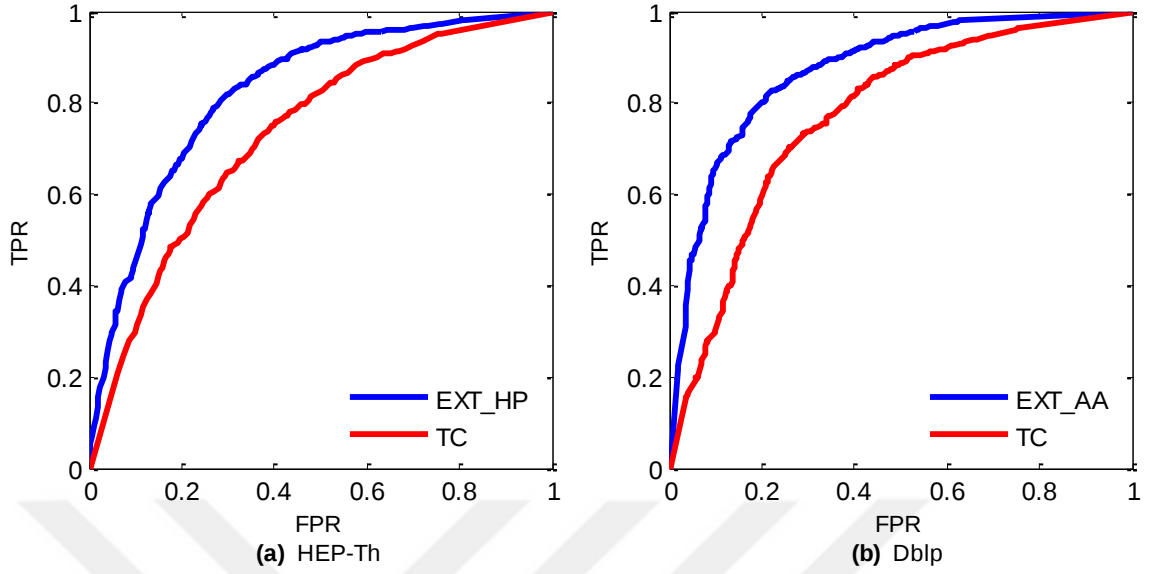
Literatürde son zamanlardaki bağlantı tahmini çalışmalarında gerçek ağların topolojik niteliklerini daha iyi kapsayan ağ modelleri kullanılarak bağlantı tahmini doğruluğunun artırılması amaçlanmıştır. Bu ağ modellerinin en popüler olanları heterojen, yönlü, ağırlıklı ve zamansal ağ modelleridir. Literatürdeki çalışmaların çoğunda, bağlantıların yön bilgisinin

Tablo 5.8. HEP-Th veri kümesinin AUROC sonuçları

IBK		RF		RT		MLP		J48	
EXT_HP	0,815	EXT_HP	0,827	EXT_RA	0,740	EXT_RA	0,816	EXT_SA	0,781
EXT_RA	0,806	EXT_RA	0,820	EXT_SA	0,740	EXT_HP	0,811	EXT_SO	0,781
EXT_SO	0,798	EXT_SA	0,819	EXT_SO	0,736	EXT_SA	0,810	EXT_HP	0,780
EXT_SA	0,797	EXT_SO	0,817	EXT_HP	0,710	EXT_SO	0,810	EXT_JC	0,774
EXT_AA	0,791	EXT_AA	0,811	TC	0,708	EXT_AA	0,808	EXT_RA	0,768
EXT_CN	0,791	EXT_CN	0,795	EXT_AA	0,688	EXT_CN	0,803	EXT_AA	0,756
EXT_JC	0,772	EXT_JC	0,790	EXT_HD	0,679	EXT_JC	0,802	EXT_CN	0,753
EXT_HD	0,768	EXT_HD	0,788	EXT_JC	0,666	EXT_HD	0,792	EXT_HD	0,752
EXT_LHN	0,756	EXT_LHN	0,781	EXT_LHN	0,656	TC	0,792	TC	0,721
TC	0,748	TC	0,735	EXT_CN	0,651	EXT_LHN	0,765	EXT_LHN	0,593

Tablo 5.9. DBLP veri kümesinin AUROC sonuçları

IBK		RF		RT		MLP		J48	
EXT_AA	0,843	EXT_AA	0,870	EXT_AA	0,822	EXT_CN	0,813	EXT_AA	0,790
EXT_CN	0,817	EXT_RA	0,847	EXT_CN	0,805	EXT_AA	0,791	EXT_CN	0,785
EXT_LHN	0,816	EXT_HP	0,836	EXT_RA	0,795	EXT_LHN	0,786	EXT_HP	0,772
EXT_HP	0,815	EXT_SO	0,830	EXT_HP	0,792	EXT_RA	0,786	EXT_LHN	0,763
EXT_RA	0,807	EXT_SA	0,828	EXT_SA	0,781	EXT_HP	0,783	EXT_JC	0,758
EXT_HD	0,788	EXT_LHN	0,827	EXT_SO	0,777	EXT_HD	0,751	EXT_RA	0,755
EXT_SO	0,786	EXT_CN	0,825	EXT_HD	0,751	EXT_JC	0,747	EXT_SA	0,750
EXT_SA	0,782	EXT_HD	0,813	EXT_LHN	0,748	EXT_SO	0,743	EXT_SO	0,750
EXT_JC	0,777	EXT_JC	0,799	TC	0,735	EXT_SA	0,716	EXT_HD	0,743
TC	0,721	TC	0,771	EXT_JC	0,731	TC	0,566	TC	0,573



Şekil 5.13. RF sınıflandırıcısında EXT_NLPM ve TC ölçüt sonuçlarının ROC eğrileri

bağlantı oluşumundaki rolü dikkate alınmamıştır. Bu bölümde komşuluk tabanlı bağlantı tahmini ölçütleri yönlü ağlarda bağlantıların yön bilgisini de hesaplama katılarak genişletilmiştir. Önerilen yöntem yönlü bağlantılardan oluşan ağ motiflerine dayalıdır. Komşuluk tabanlı ölçütlerin hesaplanması üçlü düğümlerin oluşturduğu OTP türüne göre yapılmıştır. Önerilen zamansal yaklaşımda yakın zamanda oluşan bağlantılar ağırlıklandırılmıştır. Daha fazla ağırlığa sahip olan bağlantılar, hesaplamalarda skor değerine daha fazla katkı sağlamıştır. Önerilen yöntemin bağlantı tahmini performansının doğru değerlendirilmesi amacıyla negatif örnekler eğitim verilerinde en az bir kez 2 birim uzaklıkta görülen düğüm çiftlerinden seçilmiştir. Ayrıca karmaşık ağlardaki sınıf dengesizliğinin performans değerlendirmesini etkilememesi için pozitif ve negatif örnekler eşit sayıda seçilmiştir. Bu şekilde örneklendirme yapılarak zorlu eğitim verileri oluşturulmuştur.

Deney 1'de, EXT_NLPM ölçütler ve klasik komşuluk tabanlı ölçütlerin bağlantı tahmini performansları yönlü, ağırlıklı ve zamansal atıf ağlarında karşılaştırılmıştır. İki atıf ağındaki deneysel sonuçlarda tüm EXT_NLPM ölçütlerinin komşuluk tabanlı ölçütlerden daha iyi bağlantı tahmini performansına ulaştığı görülmüştür. Bağlantı tahmini doğruluğunun daha fazla artırılması amacıyla EXT_NLPM ölçütleri ve belirlenen birkaç topolojik nitelik sınıflandırıcılarda birlikte kullanılmıştır. Elde edilen sonuçlar, esas ölçütlerin yanı sıra diğer içerik veya topolojik tabanlı niteliklerin de bağlantı tahmini performansını iyileştirme

potansiyeline sahip olduğunu göstermiştir. Deney 2'de, yönlü ağlarda bağlantı tahmini için yakın zamanda önerilen TC ölçütü EXT_NLPM ölçütleri ile karşılaştırılmıştır. TC ölçütünde bağlantıların ağırlık ve zamansal bilgileri dikkate alınmadığından adil bir karşılaştırma yapılması için karşılaştırma sadece yönlü ağlarda yapılmıştır. Deney 2'nin sonuçlarında EXT_NLPM ölçütleri büyük çoğunlukla TC ölçütünü geride bırakmıştır. Deney 1 ve Deney 2'de elde edilen sonuçlar, bağlantı oluşumunda hangi yönlü ağ motifi yapısının daha belirleyici olduğunu önerdiğimiz yöntemin etkin bir şekilde öğrenebildiğini doğrulamıştır.

Bağlantı tahmini doğruluğunun iyileştirilmesi amacıyla eğitici öğrenme algoritmalarında bu bölümde önerilen bağlantı tahmini ölçütü; yazarların makalelerinin başlık, özet vb. bilgileri, yazarların görev yaptığı enstitülerin bilgileri, makalelerin yayınlandığı dergilerin etki faktörü gibi içerik tabanlı nitelikler ve diğer potansiyel topolojik nitelikler birlikte kullanılabilir. Önerilen yönlü bağlantı tahmini yöntemi 2 birim uzaklıkla sınırlıdır. Bağlantı tahmini doğruluğunun artırılması amacıyla 2 birim uzaklıktan daha fazlasını kapsayan ağ motifleri üzerinde araştırma yapılabilir.

6. SONUÇLAR VE ÖNERİLER

Bu tezde karmaşık ağlarda bağlantı tahmini problemi için iki yöntem önerilmiştir. Birinci yöntemin amacı bilim insanlarının atıf sayısının tahmin edilmesidir. Literatürde bilimsel makalelerin atıf sayısı tahmini için çok sayıda çalışma yapılmışken bilim insanlarının atıf sayısı tahmini üzerinde yeterince durulmamıştır. Bu tezde önerilen yöntemle bu az sayıdaki çalışmalara katkıda bulunulmuştur. Bilim insanlarının atıf sayısı tahmini problemi atıf ağlarında bir bağlantı tahmini problemi olarak ele alınmıştır. Önerilen bağlantı tahmini yaklaşımında bağlantılar ağırlıklarıyla birlikte tahmin edilmiştir. Bu açıdan yapılan çalışma yönlü, ağırlıklı ve zamansal ağlarda bağlantıların ağırlıklarıyla beraber tahmin edildiği ilk çalışmadır. Atıf sayısı tahmini için atıf ağlarının zaman içerisindeki evrimi boyunca artan/azalan eğilimlerin hesaplamaya katıldığı zamansal yeni bir bağlantı tahmini ölçütü önerilmiştir. Önerilen ölçütte karmaşık ağlarda hem yerel hem de küresel topolojik yapılar bağlantı tahmini için kullanılmıştır. Testler önerilen yöntemin performansının doğru değerlendirilmesi için belirli kriterlere göre örneklendirilmiş zorlu eğitim verilerinde yapılmıştır. Bağlantı tahmini problemi kapsamında bağlantıların ağırlıklarıyla birlikte tahmin işinin zorluğu dikkate alındığında elde edilen performans sonuçları yöntemin doğruluğunu göstermiştir. Elde edilen sonuçlar bilim dünyasındaki çalışmalar arasında oluşan bibliyografik bilgilerin oluşturduğu karmaşık ağlarda veri madenciliği çalışmalarının, bilim insanlarının veya bilimsel makalelerin etki tahmini için değerli bir çaba olduğunu da ortaya koymuştur. Önerilen bağlantı tahmini ölçütü ve genel olarak iyi bir performansa sahip klasik bağlantı tahmini ölçütlerden biri olan AA ölçütüyle karmaşık ağlarda yeni bağlantı tahmini problemi kapsamında karşılaştırılmıştır. Elde edilen performans sonuçları zamansal karmaşık ağlardaki topolojik yapıların zenginliğinin ve bağlantıların zaman içerisindeki evriminin bağlantı tahmini için gelecek çalışmalara ışık tutabileceğini desteklemektedir.

Önerilen bağlantı tahmini ölçütünün farklı versiyonlarıyla elde edilen sonuçlar, ağdaki farklı topolojik yapıların ölçümelerini ayrı ayrı temsil eden bir ölçüt oluşturmanın eğitici öğrenme yöntemlerinde daha etkili olduğunu göstermiştir. Önerilen yöntemin doğruluğunun artırılması amacıyla diğer potansiyel topolojik nitelikleri ve içerik tabanlı nitelikler önerilen bağlantı tahmini ölçütüyle birlikte eğitici algoritmalarda uygulanabilirler.

Bu tez çalışmasında yapılan diğer bağlantı tahmini yönteminin amacı ise yönlü ağlarda bağlantıların yön bilgisinin bağlantı oluşumundaki rolü dikkate alınarak komşuluk tabanlı bağlantı tahmini ölçütlerinin doğruluğunun artırılmasıdır. Literatürdeki bağlantı tahmini

ölçütlerinin çoğunda bağlantıların yön bilgisinin bağlantı oluşumundaki rolü dikkate alınmamıştır. Bunun için bu çalışmada klasik komşuluk tabanlı bağlantı tahmini ölçütlerin yönlü ağ motiflerine dayalı hesaplandığı genel bir yöntem önerilmiştir. Daha etkin bir bağlantı tahmini için, önerilen yöntemde bağlantıların ağırlık ve zaman bilgileri de kullanılmıştır. Birinci yöntemde yapıldığı gibi bu yöntem de bağlantı tahmini performansının doğru değerlendirilmesi amacıyla testler zorlu eğitim veri kümelerinde yapılmıştır. Yönlü karmaşık ağlarda yapılan test sonuçları önerilen yöntemin komşuluk tabanlı bağlantı tahmini ölçütlerinin doğruluğunu dikkate değer bir şekilde artırdığı görülmüştür. Ayrıca önerilen yöntem ile bağlantıların yön bilgisini dikkate alan yakın zamanda yapılmış bir diğer çalışma olan TC ölçütüyle karşılaştırılmıştır. Elde edilen başarılı sonuçlar önerdiğimiz yöntemin TC ölçütüne göre bağlantı oluşumunda hangi yönlü ağ motifi yapısının daha belirleyici olduğunu etkin bir şekilde öğrenebildiğini doğrulamıştır.

Önerilen yönlü bağlantı tahmini yöntemi 2 birim uzaklıkla sınırlıdır. Bağlantı tahmini doğruluğunun artırılması amacıyla 2 birim uzaklıktan daha fazlasını kapsayan ağ motifleri üzerinde araştırma yapılabilir.

KAYNAKLAR

- [1] **Travers, J. ve Milgram, S.**, 1967. The small world problem, *Psychology Today*, cilt 1, no. 1, pp. 61-67.
- [2] **Backstrom, L., Boldi, P., Rosa, M., Ugander, J. ve Vigna, S.**, 2012. Four degrees of separation, *Proceedings of the 4th Annual ACM Web Science Conference*.
- [3] <https://snap.stanford.edu/data/web-Google.html> Google web graph. 1 Nisan 2018.
- [4] <https://snap.stanford.edu/data/cit-HepPh.html> High-energy physics citation network. 1 Nisan 2018.
- [5] <https://snap.stanford.edu/data/ca-AstroPh.html> Astro Physics collaboration network. 1 Nisan 2018.
- [6] <https://string-db.org/> String, Protein-Protein Interaction Networks. 1 Nisan 2018.
- [7] <http://www.disgenet.org/web/DisGeNET/menu/home> DisGeNET. 1 Nisan 2018.
- [8] **Kitisin, S. ve Neuman, C.**, 2006. Reputation-based trust-aware recommender system, *Securecomm and Workshops, 2006*.
- [9] **Walter F. E., Battiston S. ve Schweitzer F.**, 2008. A model of a trust-based recommendation system on a social network, *Autonomous Agents and Multi-Agent Systems*, cilt 16, no. 1, pp. 57-74.
- [10] **Spertus, E., Sahami, M. ve Buyukkokten, O.**, 2005. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*.
- [11] **Liben-Nowell, D. ve Kleinberg, J.**, 2007. The link prediction problem for social networks, *Journal of the American society for information science and technology*, cilt 58, no. 7, pp. 1019-1031.
- [12] **Sun, Y., Barber, R., Gupta, M., Aggarwal, C. ve Han, J.**, 2011. Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks, *2011 International Conference on Advances in Social Networks Analysis and Mining*.
- [13] **Wang, D., Pedreschi, D., Song, C., Giannotti, F. ve Barabasi, A.**, 2011. Human mobility, social ties, and link prediction, *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*.
- [14] **Dunlavy, D. K. T. G. ve Acar, E.**, 2011. Temporal link prediction using matrix and tensor factorizations, *ACM Transactions on Knowledge Discovery from Data (TKDD)*.
- [15] **Benevenuto, F., Rodrigues, T., Cha, M. ve Almeida, V.**, 2009. Characterizing user behavior in online social networks, *In Internet Measurement Conference, pages 49–62*.

- [16] **Ozgen, E. ve Baron, R.**, 2007. Social sources of information in opportunity recognition: Effects of mentors, industry networks, and professional forums, *Journal of business venturing*, cilt 22, no. 21, pp. 174-192.
- [17] **Garg, N. P., Favre, S., Salamin, H., Hakkani, D. T. ve Vinciarelli, A.**, 2008. Role recognition for meeting participants: an approach based on lexical information and social network analysis, *Proceedings of the 16th ACM international conference on Multimedia*.
- [18] **Fortunato, S.**, 2010. Community detection in graphs, *Physics Reports*, 486(3-5):75 – 174.
- [19] **Gregory, S.**, 2009. Finding overlapping communities using disjoint community detection algorithms, *In Complex Networks: CompleNet*, pages 47–61. Springer-Verlag.
- [20] **Leskovec, J., Lang, K. ve Mahoney, M. W.**, 2010. Empirical comparison of algorithms for network community detection, *In WWW*, pages 631–640.
- [21] **Qin J., Xu J. J., Hu D., Sageman M. ve Chen H.**, 2005. Analyzing terrorist networks: A case study of the global salafi jihad network, *Intelligence and security informatics*, pp. 287-304.
- [22] **Wiil, U. K., Gniadek, J. ve Memon, N.**, 2011. A Novel Method to Analyze the Importance of Links in Terrorist Networks, *Counterterrorism and Open Source Intelligence*, pp. 171-187.
- [23] **Shahmohammadi, A., Khadangi, E. ve Bagheri, A.**, 2016. Presenting new collaborative link prediction methods for activity recommendation in Facebook, *Neurocomputing*, cilt 210, pp. 217-226.
- [24] **Mori, J., Kajikawa, Y., Kashima, H. ve Sakata, I.**, 2012. Machine learning approach for finding business partners and building reciprocal relationships, *Expert Systems with Applications*, cilt 39, no. 12, pp. 10402-10407.
- [25] **Wu, S., Sun, J. ve Tang, J.**, 2013. Patent partner recommendation in enterprise social networks, *Proceedings of the sixth ACM international conference on Web search and data mining*.
- [26] **Tang, J., Wu, S., Sun, J. ve Su, H.**, 2012. Cross-domain collaboration recommendation, *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*.
- [27] **Berlusconi, G., Calderoni, F., Parolini, N., Verani, M. ve Piccardi, C.**, 2016. Link prediction in criminal networks: A tool for criminal intelligence analysis, *PloS one*, cilt 11, no. 4, p. e0154244.
- [28] **Krebs, V.**, 2002. Mapping networks of terrorist cells, *Connections*, cilt 24, no. 3, pp. 43-52.
- [29] **Zhu, J., Hong, J. ve Hughes, J.**, 2002. Using Markov models for web site link prediction, *Proceedings of the thirteenth ACM conference on Hypertext and hypermedia*.
- [30] **Sharma, P. K., Rathore, S. ve Park, J. H.**, 2017. Multilevel learning based modeling for link prediction and users' consumption preference in Online Social Networks, *Future Generation Computer Systems*.
- [31] **Li, X. ve Chen, H.**, 2009. Recommendation as link prediction: a graph kernel-based machine learning approach, *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*.

- [32] **Wang, X., Zhang, X., Zhao, C., Xie, Z., Zhang, S. ve Yi, D.,** 2015. Predicting link directions using local directed path, *Physica A: Statistical Mechanics and its Applications*, cilt 419, pp. 260-267.
- [33] **Lei, C. ve Ruan, J.,** 2012. A novel link prediction algorithm for reconstructing protein--protein interaction networks by topological similarity, *Bioinformatics*, cilt 29, no. 3, pp. 355-364.
- [34] **Davis, D., Lichtenwalter, R. ve Chawla, N.,** 2013. Supervised methods for multi-relational link prediction, *Social Network Analysis and Mining*, cilt 3, no. 2, pp. 127-141.
- [35] **Wang, P., Xu, B., Wu, Y. ve Zhou, X.,** 2015. Link prediction in social networks: the state-of-the-art, *Science China Information Sciences*, cilt 58, no. 1, pp. 1-38.
- [36] **Hasan, M. ve Zaki, M.,** 2011. A survey of link prediction in social networks, *Social network data analytics*, pp. 243-275.
- [37] **Popescul, A. ve Ungar, L.,** 2003. Statistical Relational Learning for Link Prediction, *In Proceedings of Workshop on Learning Statistical Models from Relational Data at IJCAI Conference.*
- [38] **Popescul, A. ve Ungar, L.,** 2003. Structural Logistic Regression for Link Analysis, *In Proceedings of Workshop on MultiRelational Data Mining at KDD Conference.*
- [39] **Tasker, B., Wong, M., P. Abbeel ve D. Koller,** 2003. Link Prediction in Relational Data, *NIPS'03: In Proceedings of Neural Information Processing Systems.*
- [40] **Getoor, L., Friedman, N., Koller, D. ve Taskar, B.,** 2003. Learning probabilistic models of link structure, *The Journal of Machine Learning Research*, cilt 3, pp. 679-707.
- [41] **Nallapati, R., Ahmed, A., Xing, E. ve Cohen, W.,** 2008. Joint latent topic models for text and citations, *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.*
- [42] **Airolidi, E., Blei, D., Fienberg, S., Xing, E. ve Jaakkola, T.,** 2006. Mixed membership stochastic block models for relational data with application to protein-protein interactions, *Proceedings of the international biometrics society annual meeting.*
- [43] **Xu Z., Tresp V., Yu S. ve Yu K.,** 2008. Nonparametric relational learning for social network analysis, *KDD'2008 Workshop on Social Network Mining and Analysis.*
- [44] **Leskovec, J., Kleinberg, J. ve Faloutsos, C.,** 2005. Graphs over time: densification laws, shrinking diameters and possible explanations, *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining.*
- [45] **Xiang, E. W.,** 2008. A survey on link prediction models for social network data, *PhD thesis*, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology.
- [46] **Soares, P. ve Prudencio, R.,** 2013. Proximity measures for link prediction based on temporal events, *Expert Systems with Applications*, cilt 40, no. 16, pp. 6652-6660.
- [47] **Adamic, L. A. ve Adar, E.,** 2003. Friends and neighbors on the web, *Social networks*, cilt 25, no. 3, pp. 211-230.

- [48] **Newman, M. E.**, 2001. Clustering and preferential attachment in growing networks, *Physical Review E*, cilt 64, no. 2, p. 025102.
- [49] **Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. ve Barabási, A.**, 2002. Hierarchical organization of modularity in metabolic networks, *science*, cilt 297, no. 5586, pp. 1551-1555.
- [50] **Zhou, T., L. Lü ve Zhang, Y.**, 2009. Predicting missing links via local information, *The European Physical Journal B*, cilt 71, no. 4, pp. 623--630.
- [51] **Jaccard, P.**, 1901. Etude comparative de la distribution florale dans une portion des Alpes et du Jura,.
- [52] **Leicht, E. A., Holme, P. ve Newman, M. E.**, 2006. Vertex similarity in networks, *Physical Review E*, cilt 73, no. 2, p. 026120.
- [53] **Salton, G. ve McGill, M. J.**, 1986. Introduction to modern information retrieval.
- [54] **Sørensen, T.**, 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, *Biol. Skr.*, cilt 5, pp. 1-34.
- [55] **Lü, L., Jin, C.-H. ve Zhou, T.**, 2009. Similarity index based on local paths for link prediction of complex networks, *Physical Review E*, cilt 80, no. 4, p. 046122.
- [56] **Katz, L.**, 1953. A new status index derived from sociometric analysis, *Psychometrika*, cilt 18, no. 1, pp. 39-43.
- [57] **Chen, H.-H., Gou, L., Zhang, X. L. ve Giles, C. L.**, 2012. Discovering missing links in networks using vertex similarity measures, *Proceedings of the 27th annual ACM symposium on applied computing*.
- [58] **Fouss, F., Pirotte A., Renders, J.-M. ve Saerens, M.**, 2007. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation, *IEEE Transactions on knowledge and data engineering*, cilt 19, no. 3, pp. 355-369.
- [59] **Brin, S. ve Page, L.**, 1998. The anatomy of a large-scale hypertextual Web search engine, *Computer networks and ISDN systems*, cilt 30, no. 1, pp. 107-117.
- [60] **Jeh, G. ve Widom, J.**, 2002. SimRank: a measure of structural-context similarity, *ACM*.
- [61] **Scellato, S., Noulas, A. ve Mascolo, C.**, 2011. Exploiting place features in link prediction on location-based social networks, *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*.
- [62] **Wohlfarth, T. ve Ichise, R.**, 2008. Semantic and event-based approach for link prediction, *International Conference on Practical Aspects of Knowledge Management*.
- [63] **Scripps J., Tan P.-N., Chen F. ve Esfahanian A.-H.**, 2008. A matrix alignment approach for link prediction, *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*.
- [64] **De Sá, H. R. ve Prudêncio, R. B. C.**, 2011. Supervised link prediction in weighted networks, *Neural Networks (IJCNN), The 2011 International Joint Conference on*.

- [65] **Lichtenwalter, R., Lussier, J. ve Chawla, N.,** 2010. New perspectives and methods in link prediction, *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*.
- [66] **Brouard, C., d'Alché-Buc, F. ve Szafranski, M.,** 2011. Semi-supervised penalized output kernel regression for link prediction, *28th International Conference on Machine Learning (ICML 2011)*.
- [67] **Chiang, K.-Y., Natarajan, N., Tewari, A. ve Dhillon, I. S.,** 2011. Exploiting longer cycles for link prediction in signed networks, *Proceedings of the 20th ACM international conference on Information and knowledge management*.
- [68] **Pavlov, M. ve Ichise, R.,** 2007. Finding experts by link prediction in co-authorship networks, *Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics-Volume 290*.
- [69] **Raymond, R. ve Kashima, H.,** 2010. Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs, *Joint european conference on machine learning and knowledge discovery in databases*.
- [70] **Cao, B., Liu, N. N. ve Yang, Q.,** 2010. Transfer learning for collective link prediction in multiple heterogenous domains, *Proceedings of the 27th international conference on machine learning (ICML-10)*.
- [71] **Li, X. ve Chen, H.,** 2013. Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach, *Decision Support Systems*, cilt 54, no. 2, pp. 880-890.
- [72] **Leskovec, J., Huttenlocher, D. ve Kleinberg, J.,** 2010. Predicting positive and negative links in online social networks, *Proceedings of the 19th international conference on World wide web*.
- [73] **Kunegis, J. ve Lommatzsch, A.,** 2009. Learning spectral graph transformations for link prediction, *Proceedings of the 26th Annual International Conference on Machine Learning*.
- [74] **Kashima, H., Kato, T., Yamanishi, Y., Sugiyama, M. ve Tsuda, K.,** 2009. Link propagation: A fast semi-supervised learning algorithm for link prediction, *Proceedings of the 2009 SIAM international conference on data mining*.
- [75] **Pujari, M. ve Kanawati, R.,** 2012. Link prediction in complex networks by supervised rank aggregation, *Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on*.
- [76] **Wang, C., Satuluri, V. ve Parthasarathy, S.,** 2007. Local probabilistic models for link prediction, *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*.
- [77] **Soares, P. ve Prudencio, R.,** 2012. Time series based link prediction, *Neural Networks (IJCNN), The 2012 International Joint Conference on IEEE*.
- [78] **Huang, Z. ve Lin, D. K. J.,** 2009. The time-series link prediction problem with applications in communication surveillance, *INFORMS Journal on Computing*, cilt 21, no. 2, pp. 286-303.
- [79] **Özcan, A. ve Öğüdücü, Ş. G.,** 2015. Multivariate temporal link prediction in evolving social networks, *Computer and Information Science (ICIS), 2015 IEEE/ACIS 14th International Conference on*.

- [80] **Güneş, İ., Öğüdücü, Ş. G. ve Çataltepe, Z.,** 2016. Link prediction using time series of neighborhood-based node similarity scores, *Data Mining and Knowledge Discovery*, cilt 30, no. 1, pp. 147-180.
- [81] **Huang, Q., Zhao, C., Wang, X., Zhang, X. ve Yi, D.,** 2015. Predicting the structural evolution of networks by applying multivariate time series, *Physica A: Statistical Mechanics and its Applications*, cilt 428, pp. 470-480.
- [82] **Backstrom, L. ve Leskovec, J.,** 2011. Supervised random walks: predicting and recommending links in social networks, *Proceedings of the fourth ACM international conference on Web search and data mining*.
- [83] **Menon, A. K. ve Elkan, C.,** 2011. Link prediction via matrix factorization, *Joint european conference on machine learning and knowledge discovery in databases*.
- [84] **Gao, S., Denoyer, L. ve Gallinari, P.,** 2011. Temporal link prediction by integrating content and structure information, *Proceedings of the 20th ACM international conference on Information and knowledge management*.
- [85] **Dong, Y., Tang, J., Wu, S., Tian, J., Chawla, N. V., Rao, J. ve Cao, H.,** 2012. Link prediction and recommendation across heterogeneous social networks., *In Data Mining (ICDM), 2012 IEEE 12th International Conference on*.
- [86] **Gao, S., Denoyer, L. ve Gallinari, P.,** 2011. Link pattern prediction with tensor decomposition in multi-relational networks, *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*.
- [87] **Sheng, G. A. O., Denoyer, L., Gallinari, P. ve Jun, G. U. O.,** 2012. Probabilistic latent tensor factorization model for link pattern prediction in multi-relational networks, *The Journal of China Universities of Posts and Telecommunications*, cilt 19, pp. 172-181.
- [88] **Sun, Y., Han, J., Aggarwal, C. C. ve Chawla, N. V.,** 2012. When will it happen?: relationship prediction in heterogeneous information networks, *Proceedings of the fifth ACM international conference on Web search and data mining*.
- [89] **Yang, Y., Chawla, N., Sun, Y. ve Hani, J.,** 2012. Predicting links in multi-relational and heterogeneous networks, *Data Mining (ICDM), 2012 IEEE 12th International Conference on*.
- [90] <https://kdl.cs.umass.edu/display/public/HEP-Th> HEP-Th dataset. 1 Eylül 2016.
- [91] <https://kdl.cs.umass.edu/display/public/DBLP> DBLP dataset. 1 Eylül 2016.
- [92] <http://DBLP.uni-trier.de/db/> DBLP computer science bibliography. 1 Eylül 2016.
- [93] <https://www.acm.org/> Association for Computing Machinery. 1 Nisan 2018.
- [94] **Hirsch, J. E.,** 2005. An index to quantify an individual's scientific research output, *Proceedings of the National academy of Sciences of the United States of America*, cilt 102, no. 46, p. 16569.
- [95] **Pobiedina, N. ve Ichise, R.,** 2016. Citation count prediction as a link prediction problem, *Applied Intelligence*, cilt 44, no. 2, pp. 252-268.
- [96] **Fu, L. D. ve Aliferis, C. F.,** 2010. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature, *Scientometrics*, cilt 85, no. 1, pp. 257-270.

- [97] **Yan, R., Tang, J., Liu, X., Shan, D. ve Li, X.,** 2011. Citation count prediction: learning to estimate future citations for literature, *Proceedings of the 20th ACM international conference on Information and knowledge management*.
- [98] **Didegah, F. ve Thelwall, M.,** 2013. Determinants of research citation impact in nanoscience and nanotechnology, *Journal of the Association for Information Science and Technology*, cilt 64, no. 5, pp. 1055-1064.
- [99] **Livne, A., Adar, E., Teevan, J. ve Dumais, S.,** 2013. Predicting citation counts using text and graph mining, *Proc. the iConference 2013 Workshop on Computational Scientometrics: Theory and Applications*.
- [100] **Stern, D. I.,** 2014. High-ranked social science journal articles can be identified from early citation information, *PloS one*, cilt 9, no. 11, p. e112520.
- [101] **Li, C.-T., Lin, Y.-J., Yan, R. ve Yeh, M.-Y.,** 2015. Trend-Based Citation Count Prediction for Research Articles, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.
- [102] **Dong, Y., Johnson, R. A. ve Chawla, N. V.,** 2016. Can scientific impact be predicted?, *IEEE Transactions on Big Data*, cilt 2, no. 1, pp. 18-30.
- [103] **Revesz, P. Z.,** 2014. A method for predicting citations to the scientific publications of individual researchers, *Proceedings of the 18th international database engineering & applications symposium*.
- [104] **Bütün, E., Kaya, M. ve Alhajj, R.,** 2016. A new topological metric for link prediction in directed, weighted and temporal networks, *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*.
- [105] **Jahanbakhsh, K., King, V. ve Shoja, G. C.,** 2012. Predicting missing contacts in mobile social networks, *Pervasive and Mobile Computing*, cilt 8, no. 5, pp. 698-716.
- [106] **Juszczyszyn, K., Musial, K. ve Budka, M.,** 2011. Link prediction based on subgraph evolution in dynamic social networks, *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*.
- [107] **Oyama, S., Hayashi, K. ve Kashima, H.,** 2011. Cross-temporal link prediction, *Data Mining (ICDM), 2011 IEEE 11th International Conference on*.
- [108] **Cannistraci, C. V., Alanis-Lobato, G. ve Ravasi, T.,** 2013. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks, *Scientific reports*, cilt 3.
- [109] **Liben-Nowell, D. ve Kleinberg, J.,** 2007. The link-prediction problem for social networks, *Journal of the American society for information science and technology*, cilt 58, no. 7, pp. 1019-1031.
- [110] **Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. ve Witten, I.,** 2009. The WEKA data mining software: an update, *ACM SIGKDD explorations newsletter*, cilt 11, no. 1, pp. 10-18.
- [111] **Schall, D.,** 2014. Link prediction in directed social networks, *Social Network Analysis and Mining*, cilt 4, no. 1, pp. 1-14.
- [112] **Lichtenwalter R. N. ve Chawla N. V.,** 2014. Vertex collocation profiles: theory, computation, and results, *SpringerPlus*, cilt 3, no. 1, p. 116.

- [113] **Behfar S. K., Turkina E., Cohendet P. ve Burger-Helmchen T.,** 2016. Directed networks' different link formation mechanisms causing degree distribution distinction, *Physica A: Statistical Mechanics and its Applications*, cilt 462, pp. 479-491.
- [114] **Shang K.-k., Small M. ve Yan W.-s.,** 2017. Link direction for link prediction, *Physica A: Statistical Mechanics and its Applications*, cilt 469, pp. 767-776.
- [115] **Aghabozorgi F. ve Khayyambashi M. R.,** 2018. A new similarity measure for link prediction based on local structures in social networks, *Physica A: Statistical Mechanics and its Applications*.
- [116] **Ding C. ve Li K.,** . Estimating multilateral trade behaviors on the world trade web with limited information, *Neurocomputing*, cilt 210, pp. 66-80, 2016.
- [117] **Guo F., Yang Z. ve Zhou T.,** 2013. Predicting link directions via a recursive subgraph-based ranking, *Physica A: Statistical Mechanics and its Applications*, cilt 392, no. 16, pp. 3402-3408.
- [118] **Murata T. ve Moriyasu S.,** 2007. Link prediction of social networks based on weighted proximity measures, *Proceedings of the IEEE/WIC/ACM international conference on web intelligence*.
- [119] **Brzozowski M. J. ve Romero D. M.,** 2011. Who Should I Follow? Recommending People in Directed Social Networks..

ÖZGEÇMİŞ

Ertan BÜTÜN

Elazığ – Türkiye

ertanbutun@gmail.com

KİŞİSEL BİLGİLER:

Doğum tarihi : 01.01.1984

Doğum Yeri : ELAZIĞ

EĞİTİM:

2013– devam ediyor (Doktora) Fırat Üniversitesi, Fen Bilimleri Enstitüsü,
Bilgisayar Mühendisliği, Elazığ

2011–2013 (Y.Lisans) Fırat Üniversitesi, Fen Bilimleri Enstitüsü,
Bilgisayar Mühendisliği, Elazığ

2003–2008 (Lisans) Dokuz Eylül Üniversitesi, Mühendislik Fakültesi,
Bilgisayar Mühendisliği, İzmir

2002–2003 (Dil) Dokuz Eylül Üniversitesi, Yabancı Diller Yüksek
Okulu (İngilizce Hazırlık), İzmir

İŞ TECRUBELERİ:

2011-Halen çalışıyorum Fırat Üniversitesi, Bilgisayar Mühendisliği, Araştırma
Görevlisi (2011-Halen çalışıyorum), Elazığ

2008-2010 Proklab Bilişim Sistemleri <http://proklab.com>,
Bilgisayar Mühendisi, İzmir