

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

**HOMOJEN VE HETEROJEN EVRİMSEL SOSYAL
AĞLARDA BAĞLANTI TAHMİNİ**

DOKTORA TEZİ

Alper ÖZCAN

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

EYLÜL 2017

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

**HOMOJEN VE HETEROJEN EVRİMSEL SOSYAL
AĞLARDA BAĞLANTI TAHMİNİ**

DOKTORA TEZİ

**Alper ÖZCAN
(504112502)**

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ

EYLÜL 2017

İTÜ, Fen Bilimleri Enstitüsü'nün 504112502 numaralı Doktora Öğrencisi Alper ÖZCAN, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı “HOMOJEN VE HETEROJEN EVRİMSEL SOSYAL AĞLARDA BAĞLANTI TAHMİNİ” başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı : **Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ**
İstanbul Teknik Üniversitesi

Jüri Üyeleri : **Prof. Dr. Zehra ÇATALTEPE**
İstanbul Teknik Üniversitesi

Doç. Dr. Haluk BİNGÖL
Boğaziçi Üniversitesi

Doç. Dr. Hatice KÖSE
İstanbul Teknik Üniversitesi

Doç. Dr. Ali Taylan CEMGİL
Boğaziçi Üniversitesi

Teslim Tarihi : 24 Ağustos 2017
Savunma Tarihi : 25 Eylül 2017





Aileme,



ÖNSÖZ

Öncelikli olarak doktora süresince yol gösterici fikirleri ve sağladığı motivasyon ile doktoramı tamamlamamda büyük rol oynayan, tez çalışmam süresince değerli sabrı, önerileri ve desteğiyle bu zorlu süreci bitirmemde en büyük pay sahibi olan danışmanım Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ'ye sonsuz teşekkürlerimi sunarım.

Tez ilerleme dönemlerinin tamamında tez çalışmasına sundukları değerli katkılar ve öneriler için tez jurisinde bulunan Prof. Dr. Zehra ÇATALTEPE ve Doç. Dr. Haluk Bingöl'e teşekkürü borç bilirim.

Hayatımın her aşamasında bana güvenen ve desteğini hiç esirgemeyen aileme ayrıca tez çalışmam boyunca hem sevgisini, hem de desteğini hiç esirgemeyen sevgili eşim Filiz Özlem ÖZCAN'a çok teşekkür ederim.

Doktora eğitimim süresince kurumsal desteğini esirgemeyen, doktora eğitimimi vermiş oldukları akademik izin ile yürütebilmeme imkan sağlayan Softtech AR-GE Merkezi'ne teşekkürü borç bilirim.

Ağustos 2017

Alper Özcan
(Bilgisayar Y. Mühendisi)



İÇİNDEKİLER

Sayfa

ÖNSÖZ	vii
İÇİNDEKİLER	ix
KISALTMALAR	xi
SEMBOLLER	xiii
ÇİZELGE LİSTESİ	xv
ŞEKİL LİSTESİ	xvii
ÖZET	xix
SUMMARY	xxiii
1. GİRİŞ	1
1.1 Tezin Amacı	5
2. LİTERATÜR ARAŞTIRMASI	7
2.1 Bağlantı Tahmini	7
2.1.1 Statik bağlantı tahmini yöntemleri	8
2.1.1.1 Düğüm komşuluğu tabanlı (lokal) ölçütler (topolojik)	9
2.1.1.2 Tüm yolların birleşimini temel alan (global) ölçütler (topolojik)	12
2.1.1.3 Anlamsal bilgi kullanan yöntemler	14
2.1.2 Evrimsel sosyal ağlarda bağlantı tahmini yöntemleri	15
2.1.2.1 Bağlantı tahmini için zaman serileri ve ardışık ağ serisi temsili kullanan yöntemler	15
2.1.2.2 Ağırlıklı özet ağ kullanan yöntemler	20
2.1.3 Heterojen ağlarda bağlantı tahmini	23
3. MATERYAL VE METOT	33
3.1 Karmaşık Ağlar ile İlgili Genel Kavramlar	33
3.1.1 Küçük dünya hipotezi	33
3.1.2 Demetleme katsayısı	33
3.1.3 Derece dağılımı	34
3.1.4 Rassal çizge modeli	35
3.1.5 Ölçekten bağımsız model	35
3.2 Zaman Serileri	36
3.2.1 VAR çok değişkenli zaman serisi modeli	37
3.2.2 AR, MA, ARMA / ARIMA tek değişkenli zaman serisi modelleri	39
3.3 NARX Sinir Ağı Modeli	41
3.4 Kullanılan Araç ve Yazılımlar	42
3.4.1 UJMP – Universal Java Matrix Package	42
3.4.2 Stax – Streaming XML for JAVA	42
3.4.3 R – İstatistiksel Yazılım Geliştirme Ortamı	44
4. ÖNERİLEN YÖNTEMLER	47
4.1 Kullanılan Topolojik Tabanlı Benzerlik Ölçütleri	47
4.2 Çok Değişkenli Zaman Serileri ile Bağlantı Tahmini	47
4.3 NARX ile Bağlantı Tahmini	52

5. DENEYSEL SONUÇLAR.....	55
5.1 Kullanılan Veri Kümeleri	55
5.1.1 DBLP bilgisayar bilimleri bibliyografik veri kümesi	55
5.1.2 Delicious sosyal sık kullanılanlar veri kümesi	57
5.2 Bağlantı Tahmini Test Yöntemleri	59
5.3 Test Sonuçları	60
5.3.1 VAR modeli ile homojen ağlarda bağlantı tahmini testleri	61
5.3.1.1 Farklı koşullar altında bağlantı tahmini	63
5.3.1.2 Çapraz geçerleme testleri	64
5.3.2 VAR modeli ile heterojen ağlarda bağlantı tahmini	67
5.3.2.1 Farklı bağıllık koşulları altında bağlantı tahmini	75
5.3.2.2 Çapraz geçerleme testleri	79
5.3.2.3 Farklı ölçütlerin birleştirilip eş zamanlı kullanılması	82
5.3.2.4 Düğüm derecesinin etki analizi	84
5.3.2.5 Karmaşıklık analizi	85
5.3.3 NARX modeli ile heterojen ağlarda bağlantı tahmini	86
5.3.3.1 Farklı bağıllık koşulları altında bağlantı tahmini	90
5.3.3.2 Farklı bağlantı türleri için NARX modelinin gecikme değerlerinin analizi	94
5.3.3.3 Girişler arasındaki göreceli önem dereceleri analizi	95
5.3.3.4 Karmaşıklık analizi	99
6. SONUÇ VE ÖNERİLER.....	101
KAYNAKLAR.....	103
EKLER	109
ÖZGEÇMİŞ.....	111

KISALTMALAR

AA	: Adamic Adar
AIC	: Akaike Information Criteria
AR	: Autoregressive
ARIMA	: Autoregressive Integrated Moving Average
ARIMA-Occ	: ARIMA Model using Link Occurrences
ARIMA-Sim	: ARIMA Model using Similarity Measures
ARMA	: Autoregressive Moving Average
AUC	: Area Under ROC Curve
CN	: Common Neighbor
DBLP	: Computer Science Bibliography Website
DOM	: Document Object Model
FPR	: False Positive Rate
HN	: Hidden Neurons
JC	: Jaccard Coefficient
Katz	: Katz Measure
MA	: Moving Average
MRIP	: Multi-Relational Influence Propagation
MRLP	: Multi-Relational Link Prediction
NARX	: Nonlinear Autoregressive Neural Network with External Input
NARX-Sim	: NARX Model using Similarity Measures
PA	: Preferential Attachment
PropFlow	: Propagation Flow
RA	: Resource Allocation
ROC	: Receiver Operating Characteristic
RW	: Random Walk
RWR	: Random Walk with Restart
Stax	: Streaming API for XML
TPR	: True Positive Rate
UJMP	: Universal Java Matrix Package
VAR	: Vector AutoRegression
VAR-AllSim	: VAR Model using all Similarity Measures and Link Occurrences
VAR-Sim	: VAR Model using Similarity Measures and Link Occurrences
WWW	: World Wide Web
XML	: Extensible Markup Language



SEMBOLLER

D	: NARX modeli gecikme zaman değeri
E	: Ağdaki kenarların kümesi
V	: Ağdaki düğümlerin kümesi
u,v	: Düğüm çiftlerinin temsili gösterimi





ÇİZELGE LİSTESİ

Sayfa

Çizelge 2.1 : Anlamsal yol – ilişki bilgileri.....	25
Çizelge 5.1 : DBLP veri kümesi düğüm ve bağlantı türlerine ait ilişki sayıları.....	56
Çizelge 5.2 : Delicious veri kümesi düğüm ve bağlantı türlerine ait ilişki sayıları..	58
Çizelge 5.3 : DBLP veri kümesi için, AUCall sonuçları (ağırlıksız).....	62
Çizelge 5.4 : DBLP veri kümesi için, AUCall sonuçları (ağırlıklı).....	62
Çizelge 5.5 : DBLP veri kümesi için sonuçlar.	65
Çizelge 5.6 : DBLP veri kümesi için sonuçlar.	66
Çizelge 5.7 : DBLP veri kümesi için sonuçlar.	77
Çizelge 5.8 : Delicious veri kümesi için sonuçlar.	78
Çizelge 5.9 : DBLP veri kümesi için çapraz geçerleme sonuçları.	80
Çizelge 5.10 : Delicious veri kümesi çapraz geçerleme sonuçları.	81
Çizelge 5.11 : DBLP veri kümesi bütün bağlantı türleri için farklı AUC sonuçları.	82
Çizelge 5.12 : Delicious veri kümesi bütün bağlantı türleri farklı AUC sonuçları. ..	83
Çizelge 5.13 : DBLP veri kümesi bütün bağlantı türleri için farklı AUC sonuçları.	83
Çizelge 5.14 : Delicious veri kümesi bütün bağlantı türleri farklı AUC sonuçları. ..	83
Çizelge 5.15 : DBLP veri kümesi bütün bağlantı türleri için farklı AUC sonuçları.	91
Çizelge 5.16 : Delicious veri kümesi bütün bağlantı türleri farklı AUC sonuçları. ..	92
Çizelge A.1 : Terimler ve ingilizce karşılıkları.	110



ŞEKİL LİSTESİ

Sayfa

Şekil 2.1 : Statik bağlantı tahmini yöntemlerinin aşamaları.....	8
Şekil 2.2 : Zaman skoru yöntemi (Munasinghe ve Ichise, 2012).	21
Şekil 2.3 : Heterojen bibliyografik ağ şeması (Sun ve diğ, 2011).	24
Şekil 2.4 : A-P-V-P-A anlamsal yolu (Sun ve diğ, 2011).	24
Şekil 2.5 : Örnek heterojen ağ (Davis ve diğ, 2012).	26
Şekil 2.6 : Örnek heterojen ve modifiye ağ (Lee ve diğ, 2012).	28
Şekil 2.7 : APAPA ve APA anlamsal yolu (He ve diğ, 2014).	31
Şekil 3.1 : Poisson ve Power Law derece dağılımları (Barabási, 2016).	34
Şekil 3.2 : Paralel NARX mimarisi.	43
Şekil 3.3 : Seri-Paralel NARX mimarisi.	43
Şekil 5.1 : DBLP veri kümesi yapısı.	55
Şekil 5.2 : Referans verme ve yazma bağlantı türü derece dağılımı (Url-4-KONECT, 2016).	56
Şekil 5.3 : Delicious veri kümesi yapısı.	58
Şekil 5.4 : DBLP veri kümesi “ortak yayın” bağlantı türü için, AUCall sonuçları. ..	68
Şekil 5.5 : DBLP veri kümesi “ortak atıf” bağlantı türü için, AUCall sonuçları.	68
Şekil 5.6 : DBLP veri kümesi “ortak konferans” bağlantı türü, AUCall sonuçları. ..	69
Şekil 5.7 : Delicious veri kümesi “ortak kaynak” bağlantı türü için, AUCall sonuçları.	70
Şekil 5.8 : Delicious veri kümesi “ortak etiket” bağlantı türü, AUCall sonuçları.	70
Şekil 5.9 : DBLP veri kümesi “ortak yayın” bağlantı türü için, AUCdisconn sonuçları.	72
Şekil 5.10 : DBLP veri kümesi “ortak atıf” bağlantı türü için, AUCdisconn sonuçları.	72
Şekil 5.11 : DBLP veri kümesi “ortak konferans” bağlantı türü için, AUCdisconn sonuçları.	73
Şekil 5.12 : Delicious veri kümesi “ortak kaynak” bağlantı türü için, AUCdisconn sonuçları.	74
Şekil 5.13 : Delicious veri kümesi “ortak etiket” bağlantı türü için, AUCdisconn sonuçları.	74
Şekil 5.14 : Düğüm popülerliğinin (derecesinin) etkisi.	85
Şekil 5.15 : DBLP veri kümesi “ortak yayın” bağlantı türü için, AUCall sonuçları. ..	87
Şekil 5.16 : DBLP veri kümesi “ortak atıf” bağlantı türü için, AUCall sonuçları.	87
Şekil 5.17 : DBLP veri kümesi “ortak konferans” bağlantı türü için, AUCall sonuçları.	88
Şekil 5.18 : Delicious veri kümesi “ortak kaynak” bağlantı türü için, AUCall sonuçları.	89
Şekil 5.19 : Delicious veri kümesi “ortak etiket” bağlantı türü, AUCall sonuçları. ..	89
Şekil 5.20 : Ortak-yayın bağlantı türü için gizli katmandaki nöron, gecikme ve MSE değerleri.	94

Şekil 5.21 : Ortak-kaynak bağlantı türü için gizli katmandaki nöron, gecikme ve MSE değerleri.	94
Şekil 5.22 : Ortak-kaynak bağlantı türü ve AA benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.	96
Şekil 5.23 : Ortak-atıf bağlantı türü ve AA benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.	97
Şekil 5.24 : Ortak-konferans bağlantı türü ve AA benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.....	97
Şekil 5.25 : Ortak-yayın bağlantı türü ve PropFlow benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.....	98
Şekil 5.26 : Ortak-atıf bağlantı türü ve PropFlow benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.....	98
Şekil 5.27 : Ortak-konferans bağlantı türü ve PropFlow benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.....	99

HOMOJEN VE HETEROJEN EVRİMSEL SOSYAL AĞLARDA BAĞLANTI TAHMİNİ

ÖZET

Karmaşık Evrimsel Ağlar, biyokimyasal ve ekolojik sistemlere kadar uzanan çeşitli alanlarda örnekleri olan ve zaman içinde dinamik bir şekilde topolojisi değişen ağlar olarak tanımlanmaktadır. Karmaşık ağlardaki zamana göre değişim "düğümün eklenmesi ve kaybolması", "düğüm çiftleri arasında bağlantıların eklenmesi ve kaybolması", "düğüm özelliklerinin değişmesi" olmak üzere üç farklı şekilde gerçekleşebilir. Kendi içinde karmaşık bir organizasyonu olan bu ağlarda, zaman içinde düğümlerin ve düğümler arasındaki bağlantıların değişim verisinin analizi, statik ağların analizine göre önemli bilgiler sağlamaktadır. Karmaşık ağlar doğada ve toplumda var olan çeşitli sistemleri betimlemektedir; protein ağları, İnternet, bilimsel işbirliği ağları, arkadaşlık ağları karmaşık ağlara örnek olarak verilebilir. Karmaşık Ağlar, halen çok sayıda bilinmeyen olduğu ve son yıllarda sıklıkla üzerinde çalışılan bir alandır. Karmaşık Ağların bir alt alanı olan Sosyal Ağlarda düğümler sosyal alandaki öğeleri, aralarındaki bağlantılar ise bu öğeler arasındaki etkileşimi temsil etmektedir. Sosyal ağlar homojen ve heterojen olmak üzere iki farklı türe ayrılmaktadır. Homojen ağlar, aynı tür düğümlerin bulunduğu ve bu düğümler arasındaki bağlantıların aynı türde olduğu ağlardır. Heterojen ağlar ise farklı tür düğümler ve farklı tür bağlantılar içerdiği için homojen ağlara göre çok daha zengin bilgi içerirler. Yapıları daha karmaşık olduğu için modellenmeleri homojen ağlara göre daha zordur.

Bağlantı tahmini problemi, ağ topolojisi, çeşitli benzerlik ölçütleri ve düğüm içeriği kullanılarak düğümler arası bağlantıların varlığının tahmin edilmesini amaçlamaktadır. Literatür çalışmalarında statik ve zamansal (dinamik) olmak üzere iki tip bağlantı tahmini problemi bulunmaktadır. Bağlantı tahmini probleminde yeni bağlantıların oluşup oluşmayacağı ve mevcut bağlantıların ileriki zaman dilimlerinde devam edip etmeyeceğinin tahmin edilmesi gerekmektedir. Homojen ağlarda zaman etkisini kullanan çeşitli çalışmalar yapılmış olmasına rağmen, evrimsel heterojen ağlarda zamanla değişim bilgisini kullanan yöntemler yetersiz sayıdadır. Bunun nedeni, heterojen evrimsel ağların kompleks yapısı, içermiş olduğu zengin bilgi ve zaman bilgisinin etkin kullanımının zor olmasıdır. Heterojen ağlarda, ağı oluşturan düğümlerin ve bağlantıların karşılıklı olarak sürekli bir etkileşim içinde olduğu karmaşık yapı; topolojik yapıyı oluşturan değişkenlerin tek denklemliler yerine eşanlı denklem sistemleri ile incelenmesi gereğini ortaya koymaktadır. Günümüzde homojen ve heterojen ağların zaman içinde değişimi ile ilgili çalışmalar henüz yeterli değildir. Ağların zaman içinde sürekli değişim eğiliminde olması ve düğümler arasındaki ilişkilerin çok dinamik olması nedeniyle zamanla değişim bilgisi çok önemli veriler içermektedir bu nedenle zamana göre değişim bilgisinin dikkate alınması çok önemlidir.

Tez çalışması kapsamında, homojen ve heterojen evrimsel sosyal ağlar için zaman içinde değişim analiz edilmiş ve bağlantı tahmini için aşağıda belirtilmiş olan yeni yöntemler geliştirilmiştir.

Homojen evrimsel ağlarda VAR (Vector AutoRegression) çok değişkenli zaman serisi modeli kullanılarak düğümler arasındaki bağlantıların tahmin edilmesi:

Tez çalışmasının ilk önemli katkısı, zamanla değişim bilgisi kullanılarak evrimsel homojen ağlarda düğümler arasındaki bağlantıların tahmin edilmesidir. Önerilen çok değişkenli zaman serisi modeli ile evrimsel homojen ağlarda geçmişte ortaya çıkan düğümler arasındaki benzerlik skorları ve bu düğümler arasında meydana gelen bağlantılar eş zamanlı modellenerek, gelecekte hangi düğümler arasında bağlantı meydana geleceğinin tahmin edilmesi sağlanmıştır. Zamanla ilişkili olarak tanımlanan iki veya daha fazla değişken ile ilgili yapılan ölçümleri, zamana göre sıralanmış olarak gösteren serilere “çok değişkenli zaman serisi” denir. Çok değişkenli zaman serilerine girdi olarak, düğümler arası benzerlik ölçüt skorları ve düğümler arasındaki bağlantılar kullanılmış ardından gelecekteki benzerlik ölçüt skorları tahmin edilmiştir. Yöntemin ilk aşamasında veri kümelerinin farklı ardışıl zaman dilimlerindeki görüntüleri oluşturulmuştur. Ardından, her zaman dilimindeki ağ görüntüsü için düğümler arasında çeşitli ağırlıklı ve ağırlıksız benzerlik ölçüt skorları hesaplanmıştır. Farklı zaman dilimlerinde hesaplanan benzerlik ölçüt skorları ve bağlantı sıklık bilgisi çok değişkenli zaman serisi modeli olan VAR modeline girdi olarak verilerek gelecekteki benzerlik skorları tahmin edilmiştir. Bir diğer önerilen modelde de birden fazla benzerlik ölçüt skoru VAR modeli ile eşzamanlı modellenmiş ve farklı benzerlik ölçütlerinin dinamik ilişkilerini kullanan bir model tasarlanmıştır. Ardından, gelecekteki benzerlik ölçüt skorları tahmin edilmiştir. VAR modeli, sistemde yer alan çok sayıda değişkenin geçmiş değerleri ile ifade edildiği ve her bir denklemin en küçük kareler ile çözümlendiği bir çözüm tekniğidir. Geçmiş bilginin kullanılması ile sağlanan bu yapı zaman serilerine dinamiklik katarak, birden fazla zaman serisi arasındaki ilişkileri yorumlama konusunda bir açılım yaratmıştır. Yöntemin son aşamasında, tahmin edilen benzerlik skorları kullanılarak tahmin etme başarımı, AUC cinsinden hesaplanmıştır. Önerilen modeller tek değişkenli zaman serisi modelleri ve statik bağlantı tahmini algoritmaları ile karşılaştırılmış ve daha başarılı sonuçlar verdiği görülmüştür.

Heterojen evrimsel ağlarda VAR çok değişkenli zaman serisi modeli kullanılarak düğümler arasındaki bağlantıların tahmin edilmesi:

Tez çalışmasının ikinci katkısı, zamanla değişim bilgisi kullanılarak evrimsel heterojen ağlarda düğümler arasındaki farklı türlere ait bağlantıların tahmin edilmesidir. Önerilen çok değişkenli zaman serisi olan VAR modeli ile heterojen ağın farklı zamanlardaki görüntülerinde hesaplanan çeşitli ağırlıklı ve ağırlıksız benzerlik ölçüt skorları ve bağlantı sıklık bilgisi eşzamanlı modellenmiş ve bütün bağlantı türlerine ait gelecekteki benzerlik ölçüt skorları tahmin edilmiştir. Gerçeklenen model ile zaman bilgisinin etkin şekilde kullanılması, tahmin edilecek bağlantının gözlenmiş olması kabulünün ortadan kaldırılması, düğüm çiftleri arasındaki bağlantıların ve benzerlik ölçütlerinin korelasyonunun kullanılması gerçekleşmiş ve mevcut heterojen bağlantı tahmini yöntemlerinin eksikliklerinin giderilmesi sağlanmıştır. Bağlantı tahmini için VAR modeline girdi olarak farklı bağlantı türleri için hesaplanan benzerlik skorları ve bağlantı sıklık bilgisi verilerek zaman içindeki değişimleri eşanlı ele alınmıştır. Bir diğer önerilen modelde de en iyi sonucun elde edildiği lokal (komşuluk tabanlı) ölçütler ve global (tüm yolların birleşimini temel alan) ölçütler bütün bağlantı türleri için eşzamanlı modellenmiş ve farklı benzerlik

ölçütlerin dinamik ilişkilerini kullanan bir model tasarlanmıştır. Önerilen modeller her bağlantı türünün ayrı olarak ele alındığı tek değişkenli zaman serisi modelleri ile karşılaştırılmış ve daha başarılı sonuçlar verdiği görülmüştür.

Heterojen evrimsel ağlarda NARX çok değişkenli sinir ağı modeli kullanılarak düğümler arasındaki bağlantıların tahmin edilmesi: Tezin son aşamasında, zamanla değişen heterojen sosyal ağlarda, geçmişte ortaya çıkan düğümler arasında meydana gelen farklı türlü bağlantılar ve düğümler arası benzerlik ölçütlerinin zaman içindeki değişimlerinin eşanlı kullanıldığı NARX (Doğrusal Olmayan Otoregresif Dışsal) sinir ağı modeli önerilmiş ve gerçekleşmiştir. VAR çok değişkenli zaman serisi modelinden farklı olarak NARX modeline girdi olarak farklı bağlantı türleri için hesaplanan benzerlik skorları ve bağlantı sıklık bilgisindeki değişimler verilerek eş zamanlı olarak ele alınmıştır. Girdi olarak verilen bağlantı sıklık bilgisindeki değişimler, ilgili bağlantı türü için ardışıl zaman aralıklarında iki düğüm çifti arasındaki bağlantının ağırlığının azalması durumunda "kopan bağlantı sayısı", bağlantının ağırlığının artması durumunda "oluşan bağlantı sayısı", bağlantının ağırlığının korunması durumunda "korunan bağlantı sayısı" bilgilerinden oluşmaktadır. NARX modeli iki katmanlı ileri beslemeli, döngülü ve dinamik ağlar olup zaman serisi tahminlerinde sıklıkla uygulama alanı bulmaktadır. Dinamik ağlar gecikme ve döngü mekanizmalarına sahip olmaları sebebiyle, sıralı ve zamana göre değişen veriler üzerinde kullanılabilmekte ve zaman serileri problemlerinde daha çok tercih edilmektedir. NARX modelinin bir avantajı, zaman serisi modellerinde olduğu gibi serinin durağan olmasını önşart olarak kabul etmemesi ve durağan olmayan seriler üzerinde de öngörü yapabilmesidir. Bir diğer avantajı da yine zaman serisi modellerinde olduğu gibi seriyi oluşturan veriler arasında doğrusal bir ilişkinin olduğunu varsaymaması ve hem doğrusal hem de doğrusal olmayan ilişkileri modelleyebilmesidir. Gerçeklenen modelde ayrıca NARX sinir ağındaki bağlantı ağırlıkları kullanılarak girişler arasındaki göreceli önem dereceleri heterojen ağdaki farklı bağlantı türleri için belirlenmiştir. Gerçeklenen model ile mevcut statik heterojen bağlantı tahmini yöntemlerinin eksiklikleri olan zaman bilgisinin etkin şekilde kullanılamaması ve zaman serisi kullanan modellerin eksikliği olan tahmin edilecek bağlantının gözlenmiş olması kabulünün giderilmesi sağlanmıştır.

Önerilen yöntemler iki farklı veri kümesi üzerinde test edilmiştir. Veri kümeleri, DBLP bilgisayar bilimleri bibliyografik veri kümesi ve Delicious sosyal sık kullanılanlar veri kümesinden oluşmaktadır. Veri kümeleri öğrenme ve test kümelerine ayrılmış ve öğrenme kümesi zaman serisi modellerini oluşturmak için test kümesinde modellerin başarımını ölçmek için kullanılmıştır. Dengesiz veri kümelerinde de iyi sonuç vermesi nedeniyle başarımı ölçmek için AUC kullanılmıştır. Tanımlanan çeşitli AUC ölçümleri ile yöntemlerin farklı koşullardaki başarımı ölçülmüştür. Son üç zaman diliminde bağlı olmayan düğüm çiftleri için $AUC_{disconn}$, son üç zaman diliminde bağlı olan düğüm çiftleri için AUC_{conn} , son zaman diliminde bağlı olmayan düğüm çiftleri için $AUC_{lastDisconn}$ ve son zaman diliminde bağlı olan düğüm çiftleri için de $AUC_{lastConn}$ değerleri hesaplanarak bağlantı tahmini yöntemleri değerlendirilmiştir. Kullanılan 4 farklı AUC ölçümlerine ilaveten, son zaman diliminin ardından bağlı-bağlı değil veya bağlı değil-bağlı şeklinde bağlılık durumunda değişiklik olan düğüm çiftleri için ise AUC_{change} hesaplanmıştır.

Önerilen yöntemler paralel algoritmalar tasarlanarak dağıtık ortamlarda karmaşıklığı azaltılabilir ve büyük ağlarda kullanılabilir. Ayrıca farklı tür benzerlik skorları ve farklı veri kümeleri üzerinde test edilerek genişletebilir.



LINK PREDICTION IN EVOLVING SOCIAL HOMOGENEOUS AND HETEROGENEOUS NETWORKS

SUMMARY

Complex Evolving Networks are described as networks whose topology change dynamically over time. These networks structures have examples in various areas extending from biochemical to ecological systems. Temporal changes in complex networks can actualize in three different ways including “addition and removal/disappearance of nodes”, “addition and removal of links between node pairs” and “modification of node properties/attributes”. In complex networks which have a complex organization, the analysis of variance information of nodes and links over time provides vital information compare to the analysis of the static networks. Complex networks delineate various systems existing in nature and society. Protein networks, Internet, scientific collaboration networks, and social/friendship networks can be given as examples for these type of networks. Complex Networks are an area which have still many unknowns and have been frequently studied in recent years. Nodes represent elements in social area and links among these nodes represent interaction between the elements in Social Networks which is a subfield of Complex Networks. Social networks are divided into two substantial categories: homogeneous and heterogeneous social networks. Homogeneous networks include same type of nodes and same type of links between nodes. On the other hand, heterogeneous networks consist of different type of nodes and links, and for this reason; they contain much more rich data compare to the homogeneous networks. Thus, modeling heterogeneous networks is more difficult than modeling the homogeneous networks because of their complex nature.

The problem of the link prediction aims to predict the existence probability of a links between node pairs by using network topology, various proximity (similarity) measures and node content information. In the literature, two related types of link prediction problems are considered: missing (static) and temporal (dynamic). In the case of the link prediction problem, it is necessary to estimate the prediction of both new links and repeated links. Although there are several link prediction studies in homogeneous networks using temporal information, there are very few methods available that employ time effect. The main reasons are representation difficulties of the complex structure, rich heterogeneous information, effective usage of influences between multi-typed relationships and the time effect. These characteristics of the evolving heterogeneous networks make link prediction a more challenging task. To overcome these difficulties, link prediction in such networks must model temporal evolution of the multi-typed topological features and link occurrences information of the network structure simultaneously. Consequently, link prediction in evolving heterogeneous networks must be modeled by simultaneous multivariate equations models instead of univariate models. Studies on evolving homogeneous and

heterogeneous networks do not mature to a desired level yet. Due to the dynamicity of the network structure and multi-typed topological features, the usage of network evolution for link prediction contains crucial information. Thus taking into account of the changes of topological features over time is very important.

In this thesis, we examine models for analysis of time varying information and link prediction models that combine the time information for evolving homogeneous and heterogeneous networks are implemented.

Link Prediction method in evolving homogeneous network based on a VAR (Vector Autoregression) Model for Multivariate Time Series forecasting: The first major contribution of the Thesis is the utilization of the network temporal evolution for predicting links between node pairs. The novelty of our approach lies in the method by which we apply Multivariate Time Series model to combine topological similarity measures and temporal evolutions of link occurrences information simultaneously. That enables us to predict repeated link occurrences as well as to identify new link occurrences. A multivariate time series, which involves observation and analysis of more than one statistical variable, is observed in time order. The node similarities and node connectivity information for each pair of nodes in certain time periods are utilized as input to Multivariate Time Series and future similarity measure scores are forecasted. First of all, the network is split into several time-sliced snapshots in experiments, where each snapshot represents the states of the network at different times. Then, unweighted and weighted similarity measure scores between pairs of nodes are calculated for all time-sliced snapshots. A multivariate time series model (VAR) is built for each pair of nodes in consecutive snapshots. Finally, the combination of similarity measures scores and the link occurrences information for each pair are employed as an input to the VAR model then future scores are forecasted. As a second proposed approach, the combination of all similarity measures scores and the link occurrences information between node pairs are used simultaneously and future score values are estimated. VAR is a multiple time series model in which each variable is expressed as a linear function of its own past values. The past values of the remaining variables and each equation coefficients may be estimated separately by ordinary least squares. This superior model provides a systematic way to capture rich dynamics and inter dependencies among multiple time series. In the last stage, the prediction performance in terms of AUC is evaluated on the test set by using the estimated similarity scores. The proposed models outperform the baseline static and univariate time series methods according to the comparison result.

Link Prediction method in evolving heterogeneous network based on a VAR Model for Multivariate Time Series forecasting: The second contribution of the Thesis is the proposal of a novel method to solve the problem of link prediction in evolving heterogeneous networks composed of multiple types of nodes and links that change over time and investigate the dynamics of multi-typed links in a heterogeneous social networks. The proposed model based on Multivariate Time Series (VAR) in evolving heterogeneous networks integrates the unweighted and weighted similarity measures scores and link occurrences information and future score values are estimated for all link types. Our implemented model employs the network temporal information, removes the assumption of past existence of a connection which is to be estimated and uses the correlations between multi-typed relationships and similarity measures. Thus, proposed method resolves drawbacks of the existing heterogeneous link prediction methods. The calculated node similarity

scores and node connectivity information of node pairs for multi-typed relationships are used as input to Multivariate Time Series and the variation of similarity scores and temporal evolutions of link occurrences are analyzed simultaneously. As a second proposed approach, the combination of all best local (neighborhood-based) and global (path-based) similarity measures and link occurrence information for all link types are employed simultaneously and the model using dynamic relationships of multi-typed relationships is implemented. According to the comparison of the proposed models and temporal link predictors based on univariate time series methods, the proposed approaches outperform the baseline models.

Link Prediction method in evolving heterogeneous network based on a multivariate model using the NARX Neural Network: The last contribution of the Thesis is the proposal of a novel method based on Nonlinear Autoregressive Neural Network with External Inputs (NARX) to integrate the temporal changes information for multi-typed links and the similarity measures between multi-typed links in evolving heterogeneous networks. Unlike the multivariate time series model (VAR), the calculated node similarity scores and the frequency of changes (e.g., the creation, preservation or removal of the link between node pairs) in the occurrence of links over consecutive time steps are used as input to NARX simultaneously. The frequency of changes in the occurrence of links include "Link Removal score" that demonstrates the decrease of weight value between the node pairs in the next time period, "Link Creation score" that represents the increase of weight value (increase in number of connections value) between the node pairs in the next time period and "Link Preservation score" that shows the weight value between the node pairs remaining unchanged in the consecutive time period. NARX is an implementation of recurrent dynamic networks with a two-layer feedforward network and is commonly used in time-series modeling. Due to the recurrent dynamic neural network models have feedback connections and the next value of the time series is regressed on previous values of the time series, these models are important for the forecasting of time series. Traditional time series methods require that the series to be stationary which means that the mean, variance and autocorrelation do not change over time. The other significant limitation is that traditional time series methods are only appropriate under the assumption that the time series are generated from linear processes, which is generally not accurate. NARX is a promising alternative to handle these major restrictions. NARX can reveal patterns adaptively from the time series data without prior assumption about the process from which a particular time series has generated. Hence, NARX efficiently models both linear and non-linear processes, stationary as well as non-stationary time series. Also, the proposed model determines the relative importance of each input variable based on connection weights for the corresponding link type in heterogeneous network. Our implemented model employs the network temporal information, removes the presumption of past occurrence of a connection which is to be estimated, and distinguish the nonlinear temporal evolution mechanisms of each link type simultaneously. Thus, proposed method resolves drawbacks of the existing heterogeneous link prediction methods.

Proposed methods are evaluated on two longitudinal real-world heterogeneous datasets. These data sets are; bibliographic dataset DBLP and a social bookmarking dataset Delicious. Each dataset is split into training and test sets such that training set contains time-sliced snapshots of the network that are used to create time series models and test set contains the actual links in the future. Due to the The Area Under Curve (AUC) is a commonly used performance measure in imbalanced datasets, our

link prediction models are evaluated in terms of AUC. Distinct AUC measures are defined for comparing the performance of link prediction. AUC_{disconn} is defined for only non-connected pairs in the most recent three previous time periods in the training set, AUC_{conn} is defined for only connected pairs in the most recent three previous time periods, $AUC_{\text{lastDisconn}}$ is defined for only non-connected pairs in the last time period, AUC_{lastConn} is defined for only connected pairs in the last time period. Additionally, AUC_{change} is defined for only the pairs whose state (connected or non-connected) are altered to another state after the last time period which is one of the hardest task in link prediction.

The proposed methods may be extended by designing the parallel computation methods in distributed environments to decrease the complexity and can be used in huge networks. Additionally, the proposed methods may be enhanced by using various types of node similarities and different types of network datasets.



1. GİRİŞ

Mobil teknolojilerin hızla yayılması ve çevrimiçi sosyal ağ kullanımının yaygınlaşması ile kişiler arası etkileşim de önemli ölçüde değişmiştir ve değişmeye devam etmektedir. Düğüm denen öğelerden ve bu öğeler arasındaki ilişkilerden oluşan sosyal ağlar, kendi içinde sahip olduğu karakteristik özellikleri nedeniyle aynı zamanda karmaşık ağ olarak tanımlanmaktadır. Karmaşık ağlar doğada ve toplumda var olan çeşitli karmaşık sistemleri betimlemektedir; protein ağları, bilimsel işbirliği ağları, sosyo-ekonomik ağlar, arkadaşlık ağları karmaşık ağlara örnek olarak verilebilir. Karmaşık Ağ modellerinin özellikleri incelendiğinde karmaşık ağlar ile ilgili yapılan çıkarımlarda küçük dünya hipotezi, demetleme katsayısı ve derece dağılımı olmak üzere üç kavramın önemli rol oynadığı ifade edilmektedir (Albert ve Barabasi, 2002). Bu kavramlara göre karmaşık ağların demetleme katsayılarının yüksek, ortalama en kısa yol uzunluklarının beklenenden daha kısa ve derece dağılımlarının power-law derece dağılımına uygun olduğu belirlenmiştir. Power-law derece dağılımına göre bir ağda az sayıdaki düğüm çok sayıda bağlantıya sahipken, çok sayıda düğüm az sayıda bağlantıya sahiptir. Bu özelliklere ek olarak karmaşık ağlarda zaman içinde düğümler arasındaki uzaklıklar azalmakta, düğümlerin ortalama derece sayısı ise artmaktadır (Leskovec ve diğ, 2005).

Sosyal Ağ Analizinin çeşitli alanlarda çeşitli uygulamaları bulunmaktadır. Örneğin iş hayatında bir organizasyonun içindeki bilgi/iletişim akışlarının görselleştirilmesinde, yine organizasyonlarda ekipler arasındaki iletişimin sıklığı ve kimin kiminle nasıl iletişim kurduğu gibi ilişki özelliklerinin belirlenmesinde, organizasyonlarda ekip dinamiklerinin anlaşılmasında sosyal ağ analizi kullanılabilmektedir. Sosyal ağlardaki ilişkiler, organizasyonların nasıl yönetildiğine ve bireylerin ne derece etkili olduklarının belirlenmesine kadar uzanan pek çok konuda belirleyici role sahiptir. (Burt, 2001) ağ içindeki ilişkileri sosyal sermaye olarak nitelendirmektedir. Ayrıca ağ içindeki bilgi/iletişim akışlarının zamana göre değişim bilgisi büyük önem taşımaktadır. Örneğin zaman içinde sürekli iletişim halinde olan kişiler yada ekipler arasındaki ilişki büyük ihtimal ile güçlü bir ilişkidir ve ortak yönlerin olduğunu

göstermektedir. Benzer şekilde yakın geçmişte oluşan iletişimler eski iletişimlere göre daha büyük önem taşımaktadır. Sosyal ağ analizi kritik kişilerin organizasyonda tutulması ve organizasyondaki sosyal sermayenin korunmasında önemli rol oynamaktadır örneğin güçlü bağlantıları olan kişiler durumlarından daha memnun olmaya ve dolayısıyla kurumda kalmaya daha eğilimli olmaktadır. Sosyal ağlardaki bağlantıların zayıf veya güçlü olmasının farklı anlamları bulunmaktadır. Zayıf bağlantıların gücü adı verilen kavrama göre (Granovetter, 1983, Levin ve Cross, 2004), ağ içinde bulunan gruplar (demetler) arası zayıf bağlantılar bilgi akışında daha etkin rol oynamaktadır çünkü bu zayıf bağlantılar yeni gruplara erişim sağlamaktadır. Bir diğer kavramda ağdaki iki grup arasındaki sosyal ayrılığı ifade eden yapısal boşluk kavramıdır. Aralarında yapısal boşluk bulunan iki grubu birbirine bağlayan düğümlerin sosyal sermayesi yüksektir (Burt, 2001).

Sosyal ağlar incelenirken, zamana göre ağların değişimi büyük oranda ihmal edilmiştir. Evrimsel Sosyal Ağlar olarak tanımlanan ve zaman içinde dinamik bir şekilde topolojisi değişen ağlar olarak ifade edilen Evrimsel Sosyal Ağlar'da ardışıl zaman aralıklarında, düğüm çiftleri arasında yeni bağlantıların oluşması, mevcut bağlantıların kaybolması ve düğümlerin eklenmesi ve çıkarılması şeklinde zamana göre değişimler görülmektedir. Kendi içinde karmaşık bir organizasyonu olan bu ağlarda, zaman içinde düğümlerin ve düğümler arasındaki bağlantıların değişim verisinin analizi, statik ağların analizine göre önemli veriler sağlamaktadır. Evrimsel Sosyal Ağların zamana göre nasıl değiştiği ve ağdaki değişikliğe sebep olan etmenler halen çok sayıda bilinmeyen olduğu ve son yıllarda sıklıkla üzerinde çalışılan bir alandır (Ahuja ve diğ, 2012).

Sosyal ağlar homojen ve heterojen ağlar olmak üzere iki farklı türe ayrılmaktadır. Homojen ağlar, aynı tür düğümlerin bulunduğu ve bu düğümler arasındaki bağlantıların aynı türde olduğu ağlardır. Heterojen ağlar ise farklı tür düğümler ve/veya farklı tür bağlantılar içerdiği için homojen ağlara göre çok daha zengin bilgi içerirler. Yapıları daha karmaşık olduğu için modellenmeleri homojen ağlara göre daha zordur.

Bağlantı tahmini problemi ağda iki düğüm arasındaki bağlantıların varlığını tahmin etme problemi olup, zaman içinde dinamik bir şekilde topolojisi değişen ağlarda evrimselleşmenin nedenlerini açıklayabilmektedir. Bu nedenle zamana göre değişen sosyal ağlarda hangi düğümler arasında yeni bağlantıların oluşacağını, hangi

düğümlemler arasında var olan bağlantıların korunacağını belirlenmesi sosyal ağ analizindeki en temel problemlerden biridir. Bağlantı tahmini arkadaşlık ve iş ağlarında potansiyel ilişkilerin önerilmesi, hastalık ve gen ağlarında henüz mevcut olmayan etkileşimlerin ortaya çıkarılması, büyük organizasyonlarda iş birliklerinin belirlenmesi, akademik işbirliği ağlarında ortak yayın ve ortak atıf ilişkilerinin ortaya çıkarılması, güvenlik tehditlerini ortaya çıkarmak amacıyla şüpheli kişilerin dolaylı işbirliği yaptıkları kişileri ortaya çıkarılması gibi farklı amaçlarla kullanılmaktadır. Literatürde düğümlerin bağlanma eğilimini açıklamak için önerilen sosyoloji tabanlı hipotezler şu şekilde verilmektedir: Yapısal Konumlanma (Freeman, 1979), Koheziflik (Burt, 1992), Asortatif Birleştirme (Newman, 2001b), Kümülatif Üstünlük (DiPrete ve Eirich, 2006).

Bir düğümün ağdaki diğer düğümlere doğrudan veya dolaylı olarak bağlı olması düğümün yapısal konumunu göstermektedir. Düğümlerin ağ içerisindeki yapısal konumu ağdaki bilgi akışını ve sosyal yayılım dinamiğini etkilemektedir. Freeman (1979) düğümlerin bağlanma eğilimini açıklamak için "yapısal konumlanma" hipotezini ileri sürmüştür. Bu hipoteze göre ağdaki konumları düğümleri farklılaştırmaktadır ve düğümler ağda kendilerine yakın olan düğümlerle yeni bağlantı oluşturma eğilimindedir. Örneğin bibliyografik iş birliği ağında, düğümlerin (yazarların) ağda bulundukları konumlar yazarların üretkenliklerini ve atıf alma performanslarını göstermektedir (Abbasi ve diğ, 2011, Abbasi ve diğ, 2014, Abbasi, 2016). Freeman (1979) tarafından düğümün diğer düğümlere yakınlığının derecesi olarak yakınlık (closeness) ve ağdaki önemli ve popüler düğümlerin belirlenmesi için merkezilik (centrality) ölçütleri önerilmiştir. Dört farklı merkezilik ölçütü bulunmaktadır. Derece Merkeziliği (degree centrality); bir düğümün ağdaki diğer düğümler ile arasındaki doğrudan bağlantı sayısını dikkate alarak ilgili düğümün ne kadar bağlı olduğunu ölçmektedir. Ağda hangi düğümlerin etkili ve popüler olduğunu anlamak için kullanılan bir ölçüt birimidir. Yakınlık Merkeziliği (closeness centrality); bir düğümün diğer tüm düğümlere olan en kısa yol uzunluklarının ortalaması alınarak hesaplanır ve bir düğümün sosyal ağdaki diğer düğümlere yakınlığını ölçmektedir (Freeman, 1979). Yüksek yakınlık merkeziliği değerine sahip olan düğümler, diğer düğümlere yakın yollara sahiptir. Arasındalık Merkeziliği (betweenness centrality) için ağdaki her düğümün diğer düğümlere olan en kısa yol uzunlukları hesaplanır ve her bir düğümden geçen kısa yolların sayısı arasındalık

merkeziliği deęerini verir (Freeman, 1979). Abbasi ve dię. (2012) gre arasındalık merkezilięi yksek dęmler aęda stratejik konumda bulunmaktadır, bu nedenle aęın zaman ile deęiřimi srecinde yksek olasılıkla yeni baęlantılar kuracak ve baęlı olmayan dęmlerin baęlanması aracı grevi stlenecektir. Arasındalık merkezilięi yksek olan dęmlerin belirlenmesi aędaki bilginin yayılmasında ve akışında nemli rol oynamaktadır. rneęin bir rnn satış kampanyasında kimlere reklam yapılacağının belirlenmesinde yada bir hastalığın aęda yayılmasını engellemek iin kimlere tedavi uygulanacağının belirlenmesinde arasındalık merkezilięi kullanılmaktadır. zvektr Merkezilięi (eigenvector centrality) bir dęmn aędaki dięer baęlı dęmler zerinde ne kadar doęrudan etkiye sahip olabileceğini lmektedir. Dięer merkezilik deęerlerinin hesaplanmasından farklı olarak bir dęmn zvektr merkezilięi hesaplanırken komřu dęmler hesaplamaya eřit aęırlıklı katılması yerine derece merkezilik deęeri yksek olan komřu dęmlerin etkisi daha yksek aęırlıklandırarak hesaplamaya katılmaktadır.

Koheziflik (baęlılık) hipotezine gre dęmler gemiřte baęlı olduęu dięer dęmlere gelecekte yine yksek olasılıkla baęlanma eęilimindedir (Burt, 1992). Hipotezin altında yatan temel motivasyona gre gelecekte yeni baęlantıların ortaya ıkması dęmler arasındaki gemiřteki gvene baęlıdır ve gemiřte baęlı olup birbirine gven besleyen dęm iftleri gelecekte de yksek olasılıkla tekrar baęlanacaktır. Tekrarlı baęlantılar sonucunda da dęm iftleri arasında gl baęlantılar ortaya ıkacaktır ayrıca gven dzeyinin yksek olduęu aęlarda yeni katılan dęmler aędan ayrılmayarak kalıcı olabilecektir (Levin and Cross, 2004). rneęin řirketler arasındaki iliřkiler incelendięinde gemiřte birok ticaret iliřkisi bulunan organizasyonlar gelecekte yine ticaret yapma iliřkisi kurmuřtur (Gulati, 1995). Asortatif Birleřtirme hipotezine gre dęmler gelecekte kendileriyle benzer dęm derecesine sahip dęmler ile yksek olasılıkla baęlanma eęilimindedir (Newman, 2003a, Newman, 2001c). Dęm dereceleri yakın olan dęmler arasında baęlantı kurulması daha olasıdır. Asortatif Birleřtirme hipotezinde, kendine benzer olana ilgi duyma eęilimi řeklinde tanımlanan aynı trlk (homophily) kavramı da dęmlerin hangi dęmlere baęlanacağını etkilemektedir. rneęin arkadaşlık aęlarında insanlar baęlantılarını rastgele deęil kendisine benzeyenler ve aynı kltrden olan bireyler ile kurmaktadırlar. Kmlatif stnlk hipotezine gre dęmlerin gelecekte baęlanma olasılığı dęmlerin derecesinin yksek olması ile

ilişkilidir. Buna göre yüksek dereceli düğümler gelecekte yüksek olasılıkla yeni bağlantılar edinme eğilimindedir. Sosyolojide de benzer durum Matthew etkisi şeklinde "zenginin daha zengin fakirin daha da fakir" olduğu durum olarak ifade edilmektedir (Merton ,1968).

Tez çalışması kapsamında, homojen ve heterojen evrimsel sosyal ağlar için zaman içinde değişim analiz edilmiş ve bağlantı tahmini için yeni yöntemler geliştirilmiştir. Tez çalışması, altı bölümden oluşmaktadır. Birinci bölümde tezin amacı özetlenmiştir. İkinci bölüm Literatür Araştırmasından oluşmaktadır. Üçüncü bölümde tez çalışmasında kullanılan modeller ve araçlar açıklanmıştır. Tez kapsamında önerilen ve gerçekleştirilen metodlara ilişkin detaylı bilgiler, dördüncü bölümde verilmiştir. Beşinci bölümde önerilen metodlar için yapılan testler, kullanılan veri kümeleri ve test sonuçları verilmiştir. Son bölüm olan altıncı bölümde de tez çalışmasında elde edilen sonuçlar ve gelecekte yapılması planlanan çalışmalara ait öneriler verilmiştir.

1.1 Tezin Amacı

Tez çalışmasının temel amacını, homojen ve heterojen sosyal ağlar üzerinde bağlantı tahmini problemine çözüm getirebilmek için zamanla değişim bilgisinin etkin şekilde kullanılması ve literatürde bulunan mevcut yöntemlerin başarımının artırılması oluşturmaktadır.

Tezin ilk aşamasında, evrimsel homojen ağlarda geçmişte ortaya çıkan düğümler arasında meydana gelen bağlantılar ve düğümler arası benzerlik ölçüt skorlarının değişimi eşanlı düşünülerek gelecekte hangi düğüm çiftleri arasında bağlantı meydana geleceği tahmin edilmiştir. Bağlantı tahmini için çok değişkenli zaman serileri kullanarak, bağlantı bilgisi ile benzerlik ölçütleri eşanlı zamansal analiz yapılmıştır. Bağlantı bilgisi ve benzerlik ölçütleri arasında ortaya çıkan zamansal korelasyon bilgisi ile mevcut yöntemlerin başarımı artırılmıştır.

Tezin ikinci aşamasında, evrimsel heterojen ağlarda, çok değişkenli zaman serileri ile heterojen ağı oluşturan farklı bağlantı türleri arasındaki dinamik ilişkileri ve geleceğe yönelik tahminleri açıklayabilen bir model tasarlanmıştır. Geçmişte ortaya çıkan düğümler arasında meydana gelen farklı bağlantı türleri arasındaki ilişki bilgisi ve benzerlik ölçütleri eşanlı zamansal analiz yapılarak gelecekte hangi düğümler

arasında bağlantı meydana geleceđi tahmin edilmeye alıřılmış ve mevcut yöntemlerin eksikliklerinin giderilmesi sađlanmıřtır.

Tezin üçüncü aşamasında, zaman serileri kullanılarak homojen ve heterojen ađlarda zaman bilgisi kullanılarak düđümler arasındaki bağlantıların tahmin edilmesine ek olarak, NARX yapay sinir ađları kullanılarak heterojen ađlarda zaman bilgisi kullanılarak düđüm çiftleri arasındaki bağlantı bilgisinin tahmin edilmesi sađlanmıřtır. Yapılan alıřma ile heterojen ađlarda gemiřte düđümler arasında meydana gelen bağlantı sıklık bilgisindeki deđiřim kullanılarak, gelecekte hangi düđüm çiftleri arasında bağlantı meydana geleceđinin tahmin edilmesi gereklenmiřtir.



2. LİTERATÜR ARAŞTIRMASI

2.1 Bağlantı Tahmini

Herhangi bir ağın t zamanındaki yapısı verildiğinde gelecek bir $t + \Delta t$ zaman dilimindeki, ağa yeni eklenmiş bağlantıların doğru şekilde tahmin edilmesi ve gelecekte bağlantı olup olmayacağının belirlenmesi amaçlanmaktadır (Liben-Nowell ve Kleinberg, 2003). Tahmin sonucunda ağırlıksız yönsüz ağlarda bağlantıların var olup olmadığı (var/yok), ağırlıklı yönsüz ağlarda ise bağlantının ağırlık değeri belirlenmektedir. Bağlantı tahmini arkadaşlık ve iş ağlarında potansiyel ilişkilerin önerilmesi, hastalık ve gen ağlarında henüz mevcut olmayan etkileşimlerin ortaya çıkarılması, büyük organizasyonlarda iş birliklerinin belirlenmesi, akademik işbirliği ağlarında ortak yayın, ortak atıf ilişkilerinin ortaya çıkarılması gibi farklı amaçlarla kullanılmaktadır.

Literatür çalışmalarında bağlantı tahmi üzerine yapılan çalışmalar en genel haliyle homojen ağlarda bağlantı tahmini ve heterojen ağlarda bağlantı tahmini olmak üzere ikiye ayrılmaktadır.

Homojen ağlarda yapılan literatür çalışmaları incelendiğinde topolojik özellikleri (düğüm komşuluğu ve yolların birleşimini temel alan ölçütleri kullanan yöntemler) kullanan bağlantı tahmini yöntemleri üçe ayrılmaktadır.

Birinci yöntem; zaman bilgisini kullanmadan (tek bir zaman dilimi üzerinde) düğüm benzerlik ölçütlerinin skor değerlerinin hesaplandığı statik yöntemler olarak tanımlanmaktadır. Literatürdeki ikinci yöntem ise bağlantıların sıklık bilgisinin zaman serilerinde girdi olarak kullanılmasıdır. Önerilen bu modellerde tek değişkenli zaman serileri modeli kullanılmıştır. Üçüncü yöntem ise zamanla değişen ağ bilgilerinin özeti alınarak ağırlıklı özet ağ oluşturulması ve oluşturulan ağda yakın geçmiş (recency) etkisinin (ağdaki son etkileşime ne kadar yakınsa o kadar etkili) ve etkileşim sayısının (ne kadar büyükse o kadar etkili) kullanılmasıdır.

Heterojen ağlarda yapılan literatür çalışmaları incelendiğinde ise mevcut yapılan çalışmalarda genel olarak homojen ağlar için önerilen benzerlik ölçütleri heterojen ağlara uyarlanarak bağlantı tahmini yapılmaya çalışılmaktadır.

Genel olarak literatür çalışmaları incelendiğinde statik ölçütleri kullanan yöntemlerin zaman bilgisini kullanmama ve tekrarlı bağlantıları tahmin edememe eksikliği, zaman serileri tabanlı yöntemlerin daha önce örneği görülmemiş bağlantıları tahmin edememe sorunu ve sadece zamana göre değişen tek bir değişken (farklı değişkenler arasındaki ilişki ele alınmamıştır) ile öngörü yapma eksikliği, özet ağ kullanan yöntemlerin ise yakın geçmiş etkisinin deneysel olarak bulunması ve veri kümesi bağımlı olması eksikliği tespit edilmiştir. Ayrıca heterojen ağlar için önerilen yöntemlerinde zaman bilgisini kullanmaması şeklinde belirlenen bütün bu eksikler tez çalışması kapsamında önerilen modeller ile giderilmiştir.

2.1.1 Statik bağlantı tahmini yöntemleri

Bağlantı tahmini yapılırken kullanılabilecek ağ yapısına ilişkin, topolojik özellikler (Liben-Nowell ve Kleinberg, 2003) ve nesnelerin içerik (anlamsal) özellikleri (Hsu ve diğ., 2006) olmak üzere iki çeşit veri türü mevcuttur. Literatürde önerilen modellerde, bu veri türleri kullanılmaktadır.

Topolojik bilgiyi bağlantı tahmini için kullanan statik yöntemlerde ağdaki düğüm çiftleri arasında çeşitli benzerlik ölçütlerine göre benzerlik skorları hesaplanmaktadır. Ardından hesaplanan benzerlik skoruna göre azalan bir sıralama yapılmakta ve düğüm çiftleri arasında bağlantılar tahmin edilmeye çalışılmaktadır (Şekil 2.1) (Liben-Nowell ve Kleinberg, 2003). Zaman bilgisinin kullanılmadığı (ağın tek bir zaman dilimindeki durumunda) statik ağlarda kullanılan ölçütler bu bölümde tanımlanmıştır. Ayrıca bu ölçütleri kullanan çalışmalara da yer verilmiştir. Anlamsal bilgi kullanan yöntemlerde ise düğümlere ve bağlantılara ait içerik bilgileri kullanılmaktadır.



Şekil 2.1 : Statik bağlantı tahmini yöntemlerinin aşamaları.

2.1.1.1 D ğ m komşuluđu tabanlı (lokal)  l  tler (topolojik)

Herhangi bir u ve v d ğ m  i in $\Gamma(u)$, u d ğ m n n ađ'daki komşular k mesini, $w(u, v)$ u ve v d ğ mleri arasındaki bađlantı ađırlıđını temsil etmek  zere, $\Gamma(u)$ ve $\Gamma(v)$ kesişim k mesi ne kadar b y k ise u ve v d ğ mleri o kadar y ksek olasılıkla ilerde bađlanacaktır  ng r s  buradaki temel motivasyondur. Literat rde bađlantı tahmini  alıřmalarına ilham veren Liben-Nowell ve Kleinberg (2003) tarafından ger ekleřtiren  alıřmada CN, JC, PA, AA gibi benzerlik  l  tleri ile d ğ m  iftleri arasındaki benzerlik skor deđeri hesaplanmış ve bu benzerlik skor deđeri d ğ m  iftlerinin gelecekte bađlanma olasılıđını ifade etmiřtir. Benzer řekilde literat rde farklı benzerlik  l  tleri  nerilmiş hesaplanan benzerlik deđerleri gelecekte bađlanma olasılıđını belirleyecek řekilde kullanılmıştır.

- a. Ortak Arkadařlar (Common Neighbors - CN) (Newman, 2001a): Ortak komşu sayısını kullanan bu metoda g re iki d ğ m arasındaki skor denklem 2.1'e g re hesaplanmaktadır.

$$CN(u, v) = |\Gamma(u) \cap \Gamma(v)| \quad (2.1)$$

Bađlantı ađırlıkları kullanılarak ortak arkadaşlar denklem 2.2'e g re hesaplanmaktadır.

$$WCN(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} w(u, z) + w(v, z) \quad (2.2)$$

- b. Jaccard Katsayısı (JC) (Salton ve McGill, 1983): Jaccard katsayısı, ortak komşuların toplam komşulara oranı řeklinde denklem 2.3'e g re hesaplanmaktadır.

$$JC(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \quad (2.3)$$

Bađlantı ađırlıkları kullanılarak Jaccard Katsayısı skoru denklem 2.4'e g re ađırlıklandırılmıştır (Gunes ve diđ, 2015).

$$WJC(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{w(u, z) + w(v, z)}{\sum_{a \in \Gamma(u)} w(a, u) + \sum_{b \in \Gamma(v)} w(b, v)} \quad (2.4)$$

- c. Adamic/Adar (AA) (Adamic ve Adar, 2003): Adamic/Adar  l  t nde ortak arkadaşlar temel alınır ancak seyrek g r len  zelliklere daha  ok ađırlık verilir. Diđer bir ifadeyle y ksek dereceli komşuların etkisi azaltılmıştır. Buna g re skor, denklem 2.5'e g re oluřur.

$$AA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(z)|} \quad (2.5)$$

Bağlantı ağırlıkları kullanılarak Adamic/Adar ölçüt skoru denklem 2.6'ya göre ağırlıklandırılmıştır (Barabasi ve Bonabeau, 2002).

$$WAA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{w(u,z) * w(v,z)}{\log(\sum_{c \in \Gamma(z)} w(z,c))} \quad (2.6)$$

- d. Preferential Attachment (PA) (Barabasi ve Bonabeau, 2003): Tercihli Bağlanma Katsayısı'nda komşu sayısı temel alınır ve “çok komşusu olan düğümler daha yüksek olasılıkla diğer düğümlere bağlanacaktır” motivasyonu ön plana çıkar. Hesaplama karmaşıklığı düşük bir ölçüttür. Buna göre skor, denklem 2.7'ye göre oluşur.

$$PA(u, v) = |\Gamma(u)| * |\Gamma(v)| \quad (2.7)$$

Bağlantı ağırlıkları kullanılarak Tercihli Bağlanma Katsayısı denklem 2.8'e göre ağırlıklandırılmış ve çeşitli ağlarda bağlantı tahmini için kullanılmıştır (Murata ve Moriyasu, 2007).

$$WPA(u, v) = \sum_{a \in \Gamma(u)} w(a, u) * \sum_{b \in \Gamma(v)} w(b, v) \quad (2.8)$$

- e. Resource Allocation Index (RA) (Ou ve diğ, 2007): Adamic/Adar ölçütüne benzer şekilde oluşturulmuştur. Bu ölçütte AA'ya benzer şekilde ortak arkadaşlar temel alınır ancak seyrek görülen özelliklere yüksek ağırlık verilir. Yüksek dereceli komşuların etkisi azaltılmıştır. Buna göre RA ölçütü, denklem 2.9'a göre oluşur.

$$RA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(z)|} \quad (2.9)$$

Bağlantı ağırlıkları kullanılarak denklem 2.10'a göre ağırlıklandırılmıştır (Murata ve Moriyasu, 2007).

$$WRA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{w(u,z) + w(v,z)}{\sum_{c \in \Gamma(z)} w(z,c)} \quad (2.10)$$

Zhou ve diğ. (2007) tarafından yapılan çalışmada öneri sistemlerinde RA ölçütü kullanılmıştır ayrıca Zhou ve diğ. (2009) tarafından yapılan çalışmada da RA ölçütünün CN ölçütünden başarılı sonuçlar verdiği gösterilmiştir.

- f. Cannistrai-Alanis-Ravai Index (CAR) (Cannistrai ve diğ, 2013): Düğüm çiftleri arasındaki ortak komşuluk bilgisine ek olarak ortak komşular arasındaki bağlantı sayısı da (2. seviye bağlantılar) skor hesabında

kullanılmıştır. LCL ortak komşular arasındaki bağlantı sayısını, CN ; u ve v düğümleri arasındaki ortak komşular sayısını göstermek üzere iki düğüm arasındaki benzerlik denklem 2.11’de verilmiştir.

$$CAR(u, v) = CN(u, v) * LCL(u, v) \quad (2.11)$$

Cannistrai ve diğ. (2013) kapsamında yapılan çalışma ile önerilen CAR ölçütünün farklı ağlarda CN, JC, AA, RA, PA ölçütlerinden başarılı sonuçlar verdiği gösterilmiştir.

- g. Bağlantı Tahmini için Demetleme Katsayısı (CCLP) (Wu ve diğ, 2016):

İki düğüm arasındaki benzerlik ortak komşuların demetleme katsayıları toplamı ile ifade edilmektedir. Önerilen ölçütün altında yatan temel düşünce iki düğüm arasındaki ortak komşuların demetleme katsayıları (3. Seviye bağlantılar) ne kadar yüksek ise iki düğüm gelecekte yüksek olasılıkla bağlanacaktır. Benzerlik ifadesi denklem 2.12’de verilmiştir. Denklem’de CC_z , z düğümü için lokal demetleme katsayısını ifade etmektedir. Derecesi k_z olan bir z düğümü için lokal demetleme katsayısı, L_z değeri k_z komşu arasındaki gerçekleşmiş bağlantıları göstermek üzere $C_z = \frac{2L_z}{k_z(k_z-1)}$ şeklinde ifade edilmektedir.

$$CCLP(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} CC_z \quad (2.12)$$

Wu ve diğ. (2016) yapmış oldukları çalışma ile ortak komşuların demetleme katsayılarının bağlantı tahmininde kullanılması CN ve CAR ölçütlerine göre başarımda artış sağlamıştır.

- h. Multi Relational Link Prediction (MRLP) (Davis ve diğ, 2013): Heterojen ağlar için önerilen MRLP ölçütünde, AA ölçütü heterojen ağlar için genişletilmiş ve düğüm çiftleri arasında benzerliğin bulunması için önerilen skor hesabında farklı bağlantı türlerine ait düğüm derece bilgisi denklem 2.13’teki hesaplamada dikkate alınmıştır. Denklem’de i ve j farklı bağlantı türlerini, $|\Gamma(z)|_i$ ise z düğümünün i bağlantı türüne ait derecesini göstermektedir.

$$MRLP(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log \left(\frac{|\Gamma(z)|_i}{|\Gamma(z)|_i + |\Gamma(z)|_j} \right)} \quad \begin{matrix} i=j \\ i \neq j \end{matrix} \quad (2.13)$$

MRLP ölçütünün kullanıldığı Davis ve diğ. (2011) kapsamında fenotipik ve genetik hastalıklardan oluşan heterojen ağda, bağlantı tahmini üzerine yapılan çalışmanın detayları Bölüm 2.2.3'te verilmiştir.

2.1.1.2 Tüm yolların birleşimini temel alan (global) ölçütler (topolojik)

Global ölçütlerin temel motivasyonu, iki düğüm arasında bulunan yolların sayısına göre benzerlik skoru hesaplamak veya iki düğüm arasında rassal gezinti (random walk) gerçekleştirerek hesaplanan olasılık değerine göre benzerlik skoru hesaplamaktır. Lokal ölçütlerde yakın komşulukta bulunan düğümler dikkate alınırken, global ölçütler de ilgili düğümler için ayrıca 3 adım (hop) veya daha fazla uzaklıktaki komşular da benzerlik hesaplamasına dahil edilir. Böylece ağ daha geniş bir biçimde taranır ve yakın komşuluğu bulunmayan düğüm çiftleri için benzerlik değerleri hesaplanır.

- a. Katz Index (Katz) (Katz, 1953): Katz benzerlik ölçütü iki düğüm arasındaki yolların sayısını dikkate alır, hesaplama sırasında daha kısa yollara daha yüksek ağırlık, uzun yollara ise düşük ağırlık verilir.

$$Katz(u, v) = \sum_{l=1}^{\infty} \beta^l * |paths_{u,v}^{(l)}| \quad (2.14)$$

Denklem 2.14'te; $|paths_{u,v}^{(l)}|$ u düğümünden v düğümüne " l " uzunluğunda yollar kümesini göstermektedir. β sönümleme katsayısı ise yol uzunluklarının ağırlıklandırılması için kullanılmıştır. Literatürde ifade edildiği gibi 0.05 olarak değeri belirlenmiştir.

Liben-Nowell ve Kleinberg (2003) tarafından lokal benzerlik ölçütlerine ek olarak Katz ölçütü kullanılarak çeşitli ağlar için bağlantı tahmini gerçekleştirilmiştir.

- b. Random Walk with Restart (RWR) (Tong ve diğ, 2006): RWR ölçütü, u düğümünde başlayan v düğümünde sonlanan rassal gezintinin kararlı durumdaki (steady state) olasılığı ile hesaplanır. Rassal gezinti sırasında her bir adımda α olasılık değeri ile yeniden başa dönebilir. Böylece ağın çok uzak kısımlarının gezilmesinden kurtulmuş olunur. Rassal gezinti $\alpha[0,1]$ olasılığı ile başa döner, $1-\alpha$ olasılığı ile o an bulunan düğümün komşularından rastgele birine gider. Bu işlem her adımda uygulanır ve denklem 2.15'teki gibi ifade edilir. Denklem'de $\vec{e}_u$ u . elemanı 1 diğer

değerleri 0 olan başlangıç olasılık vektörünü, $\vec{r}_u$ u düğümünün diğer düğümlere gitme olasılığını ifade eden sıralama vektörünü, r_{uv} ise u düğümünden v düğümüne ulaşmanın kararlı durumdaki olasılık değerini temsil etmektedir.

$$RWR(u, v) = r_{uv} + r_{vu} ; \quad \vec{r}_u = \alpha T \vec{r}_u + (1 - \alpha) \vec{e}_u \quad (2.15)$$

Liu ve Lü (2010) tarafından RWR ölçütü Amerikan Hava Trafiği verilerinden oluşan ağ ve Protein-Protein ağlarında bağlantı tahmini amaçlı kullanılmış ve lokal ölçütlere göre daha başarılı sonuçlar verdiği gösterilmiştir.

- c. PropFlow (PropFlow) (Lichtenwalter ve diğ., 2010): PropFlow ölçütünde kenar ağırlıkları geçiş olasılıkları olarak kullanılmakta ve kaynak düğüm ile hedef düğüm arasında rassal gezinti olasılığı ile bilgi akışı (information flow) değeri hesaplanmaktadır. Ağda gezinti için belirli bir derinlik değeri ile kısıtlı enine öncelikli arama (restricted breadth-first search) algoritması tercih edilmiştir. Algoritma genel olarak belirli bir düğümden başlayıp bir başka düğümde biten belirli uzunluktaki kısıtlı bir rassal gezintinin olasılığı ile ilgilidir. Bir sonraki düğümün belirlenmesi açısından “PageRank” algoritmasına (Brin ve Page, 1998) benzer bir algoritma tercih edilmekte fakat “PageRank” algoritması gibi gezintinin tekrar başlatılmasına veya yakınsamasına ihtiyaç duyulmamaktadır. Belirli bir derinlik değeri ile kısıtlı enine öncelikli arama algoritması tercih edildiği için önerilen yöntemde rassal gezinti ile kaynak düğümden hedef düğüme derinlik değeri kısıtı altında ulaşılmaya çalışılmaktadır. Rassal gezintide düğümler tekrar edilmemektedir. Denklem 2.16’da $NodeInput_u$ değeri u düğümünün giriş olasılık değerini, w_{uv} u ve v düğümleri arasındaki bağlantı ağırlığını, $N(u)$ ise derinliği u düğümünün derinliğinden büyük olan u düğümünün komşularını temsil etmektedir. Derinlik değeri başlangıç düğümü ile ilgili düğüm arasındaki yolda bulunan düğüm sayısını ifade etmektedir.

$$PropFlow(u, v) = NodeInput_u * \frac{w_{uv}}{\sum_{k \in N(u)} w_{uk}} \quad (2.16)$$

Leichtwalter ve diğ. (2010) tarafından gerçekleştirilen çalışmada düğüm çiftleri için PropFlow benzerlik ölçüt skorları hesaplanmakta ardından hesaplanan benzerlik skoruna göre azalan bir sıralama yapılarak düğüm çiftleri arasında bağlantılar tahmin edilmeye çalışılmaktadır. PropFlow ölçütüne ek olarak gözetimli (supervised) yöntem önerilmiştir ve bağlantıların

olup olmadığı belirlenmiştir. Önerilen gözetimli yöntemde veri kümesinin hazırlanması aşamasında veri kümesi ardışık iki zaman aralığına bölünmüştür. İlk zaman aralığında topolojik özellikler ve düğüm özellikleri ele alınırken ikinci zaman aralığında ise bağlantı olması durumunda pozitif, bağlantı olmaması durumunda ise negatif sınıf etiketleri belirlenir. Sınıflandırma yöntemi olarak WEKA kütüphanesinde bulunan C4.5, Naive Bayes ve WEKA bagging yöntemleri tercih edilmiştir. Sınıflandırıcı başarımları ROC eğrisi altında kalan alana (AUC) göre karşılaştırılmıştır. Önerilen gözetimli yöntemin gözetimsiz yöntemden %15 daha yüksek başarı sağladığı iddia edilmektedir.

- d. Multi-Relational Influence Propagation (MRIP) (Yang ve diğ., 2012) Heterojen ağlar için önerilen MRIP ölçütünde K farklı bağlantı türü sayısını, $\sigma(i, j)$ bağlantı türü i ve j 'nin koşullu olasılığını göstermek olmak üzere i bağlantı türü için u ve v düğümleri arasındaki benzerlik ölçüt skoru denklem 2.17'deki gibi bulunur. Denklem'de $w(u, v)_i$ u ve v düğümleri arasındaki i bağlantı türüne ait bağlantıların ağırlığını, $|\Gamma(u)|_i$ ise u düğümünün i bağlantı türüne ait derecesini göstermektedir.

$$MRIP_i(u, v) = NodeInput_u * \beta * \frac{w(u, v)_i}{|\Gamma(u)|_i} + NodeInput_u * \beta * \sum_{j \neq i}^K \frac{\sigma(i, j) * \frac{w(u, v)_j}{|\Gamma(u)|_j}}{K-1} \quad (2.17)$$

MRIP ölçütünün kullanıldığı Yang ve diğ. (2012) tarafından yapılan çalışmada aynı tür düğümler ve farklı tür ilişkilerden oluşan bibliyografik ve protein-protein ağlarında kaynak düğüm ve hedef düğüm çiftleri arasında MRIP ölçütüne göre benzerlik skorları hesaplanmış hesaplanan skorlar bağlantı tahmini amaçlı kullanılarak PropFlow ölçütünden başarılı sonuçlar verdiği gösterilmiştir.

2.1.1.3 Anlamsal bilgi kullanan yöntemler

Hsu ve diğ. (2006) tarafından ağın bağlantı yapısı ile içerik (anlamsal) bilgisini birlikte kullanan bağlantı tahmini yöntemi önermiştir. Çalışmanın motivasyonu, kişilere yeni arkadaş önerilerinde bulunurken, içerik bilgisinden elde edilen ortak ilgi bilgisine ek olarak, ağın topolojik yapısını da kullanmanın bağlantı tahmini başarımını arttırmasıdır. Çalışma için LiveJournal sitesinden toplanan arkadaşlık

bilgileri kullanılmıştır. Sonuç olarak ağ yapısı ve düğüm içerik bilgilerini birlikte kullanan yöntemin, sadece düğümlerin içerik bilgisini kullanan yöntemle göre başarılı olduğu görülmüştür.

2.1.2 Evrimsel sosyal ağlarda bağlantı tahmini yöntemleri

Bu bölümde, zamanla değişen ağlarda gelecekte oluşabilecek bağlantıları tahmin etmek için literatürde bulunan yöntemlere yer verilmiştir.

2.1.2.1 Bağlantı tahmini için zaman serileri ve ardışık ağ serisi temsili kullanan yöntemler

Evrimsel ağlarda zaman içinde ağ değişim bilgisi kritik olduğu için ağdaki bağlantıların tahmini için ağın mevcut statik görüntüsünün kullanımı yerine ağın zaman içinde değişen görüntüleri kullanılmalıdır. Ayrıca ağda zaman içinde tekrarlanan bağlantıların ortaya çıkarılması için de zamanla değişim verisinin kullanılması gerekmektedir. Statik yöntemlerin yetersiz kaldığı bu gibi durumlar için Huan ve Lin (2009) tarafından bağlantıların meydana gelme sıklığının tek değişkenli (univariate) zaman serileri ile modellenmesi önerilmiştir. Çalışmada sadece bağlantı yapısı ve ağın topolojisi kullanılmaktadır ve her düğüm çifti arasındaki bağlantılar ayrı bir zaman serisi olarak modellenmektedir. Önerilen yöntemde, bağlantılar arası bir ilişkinin olmadığı varsayılmakta ve gelecek zaman dilimindeki bağlantılar, sadece ilgili geçmiş bağlantı bilgisine göre belirlenmektedir. Test aşamasında literatürdeki bağlantı tahmini yöntemleri ile benzer sonuçlar elde edilmiştir. Önerilen bu modele ilaveten, model mevcut statik yöntemlerle birleştirilerek hibrit bir model elde edilmiştir. Hibrit modelde mevcut statik yöntemlere göre daha iyi sonuç elde edilmiştir.

Çalışmada tek değişkenli zaman serisi yöntemi olarak ARIMA (Autoregressive Integrated Moving Average) Modeli (Box ve Jenkins 1970) tercih edilmiştir. ARIMA modeli, AR (Autoregressive) ve MA (Moving Average) modellerini içermektedir. Ayrıca ARIMA modeli; istatistiksel proses, finansal öngörü, trafik yoğunluğu simülasyonu gibi uygulamalarda sıklıkla kullanılan modellerden biridir. Çalışmada statik bağlantı tahmin yöntemleri ve statik topolojik ölçütlerinden yararlanılmıştır. İlgili ölçütler; ortak komşu sayısına dayanan CN, düğüm derecelerinin çarpımını kullanan PA, her ortak komşuya düğüm derecesi ile ters orantılı bir ağırlık belirleyen AA, düğüm çiftleri arasındaki yolların kullanıldığı Katz

ölçütlerinden oluşmaktadır. Gelecek dönemde planlanan çalışma olarak çok değişkenli zaman serileri analizinin kullanılması planlanmıştır.

Çalışmada önerilen yöntem, ENRON e-posta veri kümesi ve HEP-Th yüksek enerjili fizik makalelerinden oluşan bibliyografik veri kümesinde test edilmiştir. Bağlantı tahmini başarımını değerlendirmek için AUC yöntemi kullanılmıştır. Testlerden elde edilen sonuçlara göre, tek değişkenli zaman serileri yönteminin, bazı statik yöntemlerden daha başarılı sonuçlar verdiği fakat hibrit yöntem kullanıldığında statik yöntemlerin tümünden daha başarılı performans elde edildiği belirtilmiştir. Özellikle tekrarlı bağlantıların yoğun olduğu veri kümelerinde başarımı yüksek sonuçlar elde edilmiştir. Ayrıca tek değişkenli zaman serileri ile başarımın artırılması, çok değişkenli zaman serileri (multivariate time series) için de motive edici olmuştur.

Bir diğer çalışmada (Soares ve Prudencio, 2012) modelleme için zaman ilişkisi önemli olduğundan, zaman serileri tercih edilmiştir. Tek değişkenli zaman serisi yöntemleri olarak; AR Süreci, MA Süreci, Rassal Yürüyüş Süreci, Üstel Düzleştirme (Exponential smoothing) modelleri seçilmiştir. AR Süreci'nde (denklem 2.18) eşitliğin sağ tarafı bağımlı değişkenin gecikmeli değerlerinden oluşur. Hata terimi ise rassal sürece sahiptir ve sıfır ortalama ve sabit varyans değerine sahiptir.

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_p y_{t-p} + e_t \quad (2.18)$$

Yukarıdaki eşitlik sabit parametrenin olmadığı ve bağımlı değişkenin değerinin sadece geçmiş değerleri ve hata terimi ile açıklandığı modeldir. Bu sürece AR(p) süreci denilmektedir. Gecikme değeri p 'nin ne olacağına ilişkin karar minimum Akaike Bilgi Kriteri (AIC) değeri ile belirlenir. AIC farklı modeller arasından en optimumunu seçmek ve modeldeki değişken sayısını belirlemek için kullanılmaktadır. Mevcut modeller arasında AIC değerinin en küçük olduğu model optimum model olarak seçilir ve böylece az sayıda değişken ile yüksek başarımlar sağlanır (Akaike, 1973).

MA Süreci'nde (denklem 2.19) bağımlı değişken, beyaz gürültü sürecine sahip geçmiş hata terimleri ile ifade edilmektedir.

$$y_t = \beta_0 + \beta_1 e_t + \beta_2 e_{t-1} + \beta_3 e_{t-1} + \dots + \beta_q e_{t-q} \quad (2.19)$$

Bu eşitlikte eşitliğin sağ tarafı anlamlı q sayısı kadar hata terimi ile ifade edilmektedir. Hata terimi, sıfır ortalamaya ve sabit varyansa sahiptir.

Rassal Yürüyüş Süreci (denklem 2.20), stokastik sürecin en basit formudur. Bir serinin zaman içindeki değişimi, rassal sebeplere bağlı ise yani düzenli bir hareket sergilemiyorsa rassal yürüyüş modelinden söz edilebilmektedir.

$$y_t = y_{t-1} + e_t \quad (2.20)$$

Üstel Düzleştirme (denklem 2.21), serideki mevsimsel ve rassal etkileri dikkate alarak serinin genel eğilimini ortaya koymaya çalışan ve bu yolla geleceği tahmin edebilen yöntemdir.

$$y = \alpha y + (1 - \alpha)y_{t-1} \quad (2.21)$$

Bu çalışmada bir de gözetimli (supervised) yöntem önerilmiştir. Sınıflandırıcı olarak WEKA karar destek makineleri algoritması yöntemi tercih edilmiştir. Sınıflandırıcı için nitelik kümesi düğüm çiftleri arasındaki hesaplanmış olan topolojik statik ölçütlerden (AA, CN, JC, PA) oluşan skorlardan oluşmaktadır. Çalışma için gerekli veri, HEP-Th bibliyografik ağ sitesinden toplanmakta ve 1990 ile 2000 yılları arasında yüksek enerjili fizik alanında yayınlanan makalelere ilişkin bilgileri içermektedir. Bağlantı tahmini başarımı için AUC değeri kullanılmıştır. Çalışmada yazarlar arasındaki “birlikte yayın yapma” ilişkileri tahmin edilmektedir. Testler sonucunda, zaman serileri yöntemlerinin ve gözetimli yöntemin statik yöntemlerden daha iyi sonuç verdiği görülmüştür.

Gunes ve diğ. (2015) tarafından önerilen TS (Time Series of Similarities) yönteminde Soares ve Prudencio (2012)’da önerilen tek değişkenli zaman serileri modellerinden farklı olarak ARIMA modeli kullanılmıştır. Hesaplanan benzerlik ölçüt skorları ARIMA modeli ile analiz edilmiş ve evrimsel homojen ağlarda düğüm çiftleri arasındaki gelecek benzerlik skor değerleri hesaplanmıştır. Önerilen modelde AR, MA, RW modellerine göre daha başarılı sonuçlar elde edilmiştir.

Soares ve Prudencio (2013) tarafından önerilen olay tabanlı yöntemde, veri kümesi belirli zaman dilimlerine bölünerek zamanla değişen çok sayıda ağ görüntüsü (ardışık ağ dizisi ("snapshots")) oluşturulmakta ve ardışık iki zaman aralığında bağlantılar arasında zamansal olaylar ("temporal events") tanımlanarak yeni bir benzerlik ölçütü önerilmektedir.

E_{k-1} ve E_k kümeleri u ve v düğüm çifti arasında bulunan ardışık iki zaman aralığındaki bağlantı kümelerini temsil etmek üzere; zamansal olaylara bağlı benzerlik skorları şu şekilde hesaplanmaktadır.

u ve v bağılı düğüm çifti arasındaki bağlantının kopması durumunda gerileyen (“regressive”) bir zamansal olay gerçekleşmekte (denklem 2.22) pozitif olmayan sabit r skoru atanmaktadır.

$$R(u, v, k) = \begin{cases} r, (u, v) \in E_{k-1} \setminus E_k \\ 0, \text{diğer durumlar} \end{cases} \quad (2.22)$$

Bağılı olmayan u ve v düğüm çiftleri arasında bağlantının oluşması durumunda gelişen (“innovative”) (denklem 2.23) bir zamansal olay gerçekleşmekte ve pozitif sabit bir i skoru atanmaktadır.

$$I(u, v, k) = \begin{cases} i, (u, v) \in E_k \setminus E_{k-1} \\ 0, \text{diğer durumlar} \end{cases} \quad (2.23)$$

Bağılı olan u ve v düğüm çifti arasında bağlantının korunması durumunda ise durağan (“conservative”) (denklem 2.24) bir zamansal olay gerçekleşmekte ve pozitif sabit bir c skoru atanmaktadır.

$$C(u, v, k) = \begin{cases} c, (u, v) \in E_{k-1} \cap E_k \\ 0, \text{diğer durumlar} \end{cases} \quad (2.24)$$

Çalışmada ardışık iki zaman aralığında bağlantılar arası gerçekleşen zamansal olaylara karşılık atanan r , i , c skorlarının katsayıları deneysel olarak belirlenmiştir. Bu nedenle çalışmada önerilen benzerlik skoru veri kümesi bağımlıdır. Çalışmada önerilen iki düğüm arasındaki benzerliği ifade eden toplam benzerlik skoru birincil zamansal olaylar (“primary temporal events”) ve ikincil zamansal olaylar (“secondary temporal events”) olmak üzere iki bileşenden oluşmaktadır. n toplam zaman dilimi sayısını, α ikincil olayların katkı katsayısını göstermek üzere, u ve v düğüm çifti arasındaki benzerlik skoru denklem 2.25’deki gibi hesaplanmaktadır.

$$skor(u, v) = \sum_{k=2}^n P(u, v, k) + \alpha * S(u, v, k) \quad (2.25)$$

Birincil zamansal olaylar bileşeni $k - 1$ ve k ardışık zaman dilimleri olmak üzere u, v düğüm çifti arasındaki gerçekleşen zamansal olayların skorlarının toplamından oluşmaktadır (denklem 2.26).

$$P(u, v, k) = C(u, v, k) + I(u, v, k) + R(u, v, k) \quad (2.26)$$

u, v düğüm çifti arasındaki ikincil zamansal olaylar bileşeni ise u, v düğümleri ile u, v düğümlerinin ortak komşuları arasında gerçekleşen birincil zamansal olayların toplamından oluşmaktadır (denklem 2.27).

$$S(u, v, k) = \sum_{y \in \Gamma(u) \cap \Gamma(v)} P(u, y, k) + P(y, v, k) \quad (2.27)$$

Çalışmada ayrıca yakın geçmiş etkisini modellemek için $\beta(k) = \log(k)$ olmak üzere skor değeri β katsayısı ile çarpılmıştır (denklem 2.28). Skor hesabında bu katsayının kullanılması ile yakın geçmiş etkisinin, eski bağlantılara göre daha yüksek oranda ağırlıklandırılması amaçlanmaktadır. Böylece düğüm çiftleri arasında zaman etkisinin de kullanıldığı bir hesaplama gerçekleştirilmiş olunur.

$$skor(u, v) = \sum_{k=2}^n \beta(k) [P(u, v, k) + \alpha * S(u, v, k)] \quad (2.28)$$

Önerilen modelin test edilmesi için yapılan deneysel çalışmada veri kümesi olarak 1992-2009 yılları arasındaki 18 zaman dilimini kapsayan yüksek enerjili fizik makalelerinden oluşan HEP-Th veri kümesi kullanılmıştır. Bağlantı tahmini sonucunun ölçülmesi için AUC kullanılmıştır. Testler sonucunda çalışmada önerilen zamansal olay (“event-based”) tabanlı yöntemin statik yöntemlerden (AA, CN, JC, PA) daha iyi sonuç verdiği görülmüştür.

Farklı zaman dilimlerindeki benzerlik ölçüt skorlarını ve zaman etkisini kullanan bir başka yöntem de Moradabadi ve Meybodi (2016) tarafından önerilen “CALA-LP (Continuous Action set Learning Automata)” yöntemidir. Önerilen yöntemde homojen ağlar için bağlantı tahmini problemi optimizasyon problemi olarak ele alınmış ve Pekiştirmeli Öğrenme yöntemi kullanılmıştır. Yöntem, $q(s, a)$ olarak gösterilen durum-eylem değerinin alınan geri bildirimlere göre güncellenmesini sağlar. Bu şekilde değer (fayda) fonksiyonunu maksimize eden optimum parametrelerin belirlenmesi sağlanır. $Skor(i) = \sum_{t=1}^{T+1} \gamma^{T-t+1} (x_t(i) W_t(i) - \theta_t)$, $T+1$ 'ındaki benzerlik skorunu ifade etmek üzere, $x_t(i) = \{x_{t1}, x_{t2}, \dots, x_{tn}\}$, i . bağ için t zamanındaki nitelik (feature) vektörünü göstermektedir.

Burada her bir nitelik farklı bir benzerlik ölçütü için benzerlik skorunu ifade etmektedir. $W_t(i)$, ilgili benzerlik ölçüt skorunun ağırlık katsayısını, γ zaman etki katsayısını, θ_t eşik değerini göstermektedir. Önerilen yöntem, skor fonksiyonunu maksimize eden parametreleri belirlemekte ve sonuç olarak elde ettiği skor değerini bağlantı tahmini için kullanmaktadır. Önerilen yöntem farklı zaman dilimlerindeki farklı benzerlik ölçütlerinin, nihai skor üzerine etkilerini farklı katsayılar ile hesaba katarak bağlantı tahmini yapmaktadır. Diğer bir ifade ile hesaplanan skor değeri farklı benzerlik ölçütlerinin kombinasyonundan oluşmaktadır.

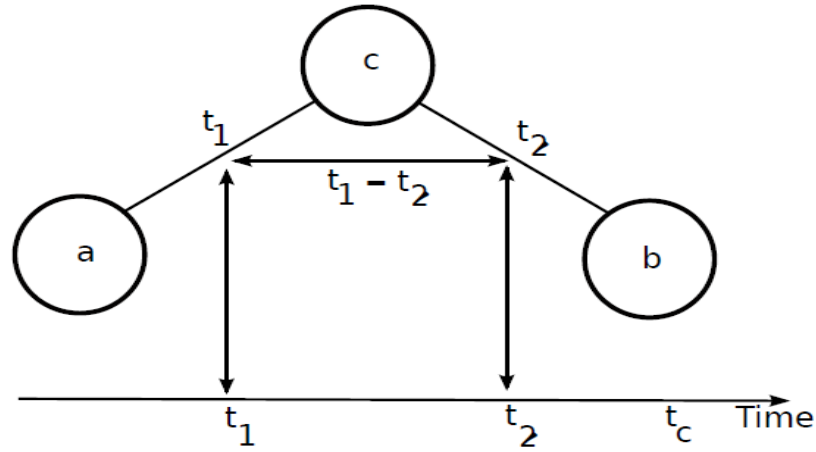
Yao ve diğ. (2016) kapsamında, mevcut yöntemler genişletilerek dinamik olarak değişen evrimsel homojen ağlar için bağlantılar tahmin edilmeye çalışılmıştır. Önerilen model ortak komşular ölçütünü genişleterek “zamana göre değişen ağırlıklı benzerlik ölçütü” ve “ortak komşuların bağlılığı ölçütü” olmak üzere iki farklı benzerlik ölçütü önerilmiştir. “zamana göre değişen ağırlıklı benzerlik ölçütü” ile yakın geçmiş etkisi, “ortak komşuların bağlılığı ölçütü” ile ortak komşuların benzerliği kullanılarak iki düğümün benzerliğinin modellenmesi sağlanmıştır. Model DBLP bibliyografik veri kümesi üzerinde test edilmiştir. Çalışmada CN ölçütünü genişleterek önerilen ölçütlerin PA, RA, CN ölçütlerinden daha başarılı sonuçlar verdiği gösterilmiştir.

2.1.2.2 Ağırlıklı özet ağ kullanan yöntemler

Bir diğer yapılan çalışmada “Time Score (TS)” ölçütü önerilmiş (Munasinghe ve Ichise, 2012) ve aday olarak belirlenen 2 düğüm çifti arasındaki benzerliği hesaplamak için ortak komşulara olan bağlantıların ağırlıklarının harmonik ortalaması ve yakın geçmiş etkisi değeri çarpılarak benzerlik skoru hesaplanmıştır (denklem 2.29).

$$TS(a, b) = \sum_n \frac{H_{m_n}(1-\alpha)^{k_n}}{|t_{1_n} - t_{2_n}| + 1} \quad (2.29)$$

Yakın geçmiş etkisi büyük olan ve ağırlığı fazla olan bağlantılara sahip olan ortak komşular daha etkili olmaktadır.



Şekil 2.2 : Zaman skoru yöntemi (Munasinghe ve Ichise, 2012).

Önerilen zaman skoru ölçütünde yakın geçmiş etkisi ve zamanla değişim bilgisini modellemek için n ortak arkadaş sayısı olmak üzere düğüm çiftleri arasındaki benzerlik skoru denklem 2.28'deki formül ile hesaplanmaktadır. Şekil 2.2'ye göre, c düğümü a ve b düğümlerinin ortak komşusunu temsil etmekte ve t_1 , a ile c düğümü arasındaki en son etkileşim zamanını göstermektedir. Benzer şekilde t_2 , b ile c düğümü arasındaki en son etkileşim zamanını göstermektedir. H_m değeri ise bağlantı ağırlıklarının harmonik ortalamasıdır. k değeri ise t_c şimdiki zaman olmak üzere $k = t_c - \max(t_1, t_2)$ şeklinde hesaplanmaktadır.

Önerilen yöntemde yakın geçmiş etkisi ve zamanla değişim bilgisini modellemek için $(1 - \alpha)^{k_n}$, α değeri 0 ile 1 arasında olmak üzere zamanla yakın geçmiş etkisi azalma katsayısı kullanılmıştır. Skor hesabında bu katsayının kullanılması ile yakın geçmiş etkisinin daha yüksek önem taşıması gerçekleşmiştir. Eğer bir düğüm çifti yakın zamanda etkileşim içinde ise düğüm çifti arasındaki bağlantının yakın geçmiş etkisi fazladır ve düğüm çiftleri arasındaki yakın geçmiş etkisi değeri zamana göre azalmaktadır. Deneysel çalışmalarda α değeri 0.1 olarak belirlenmiştir.

Bu çalışmada gözetimli yöntem önerilmiştir. Sınıflandırıcı olarak WEKA C4.5 karar ağacı algoritması yöntemi tercih edilmiştir. Önerilen modelin başarımı göstermek için yapılan deneysel çalışmada veri kümesi olarak 2007-2009 yılları arasındaki 60.000 kullanıcıdan oluşan facebook kullanıcı ağı ve 2 farklı nitelik kümesi kullanılmıştır. Birinci nitelik kümesi düğüm çiftleri arasındaki hesaplanmış olan statik topolojik ölçütlerden (AA, CN, JC, PA) oluşan niteliklerden, ikinci küme ise topolojik statik ölçütler ve önerilen zaman skoru değerlerinden oluşmaktadır. Veri kümesi eğitim ve test kümesi olmak üzere ikiye ayrılmış ve C4.5 karar ağacı

algoritması kullanılarak gözetimli öğrenme yöntemi ile önerilen yöntemin başarımı ölçülmüştür. Başarımının statik ölçütlere göre daha yüksek olduğu F-ölçütü değeri ile gösterilmiştir. Yakın geçmiş etkisinin kullanıldığı önerilen yöntemde statik ölçütlere göre F-ölçütü değerinde %4 oranında daha fazla yüksek başarımla sağladığı iddia edilmektedir.

Munasinghe ve Ichise (2012) ayrıca yakın geçmiş etkisi ve zamanla değişim bilgisini modellemek için düğümler arasındaki kenarlar zaman bilgisi ile etiketlenerek PropFlow ölçütü genişletilerek T_Flow ölçütü önerilmiştir. Birbirlerinden rastgele uzaklıkta olan düğüm çiftleri arasında bağlantı tahmini yapmak için ilgili düğüm çiftleri arasındaki en kısa yollar ele alınmakta ve bu yollar üzerinde her düğüm çifti arasındaki son etkileşim zamanı ve bağlantı ağırlıkları kullanılmaktadır.

$$T_{Flow(i,j)} = NodeInput_i * \frac{w_{ij}}{\sum_{k \in N(i)} w_{ik}} * (1 - \alpha^{|t_x - t_y|}) \quad (2.30)$$

w_{ij} , i kaynak düğümü ile j hedef düğümü arasındaki kenar ağırlığını, $N(i)$ derinliği i düğümünün derinlik değerinden büyük i düğümünün komşularını göstermektedir (denklem 2.30). $NodeInput_i$ değeri ise i düğümüne gelen olasılık değeridir ve başlangıç düğümü için değeri 1'dir. t_x değeri i düğümüne gelen bağlantının zaman etiketini göstermekte t_y ise i ve j düğümü arasındaki zaman etiketi değeridir. α değeri 0 ile 1 arasında olmak üzere zamanla yakın geçmiş etkisi azalma katsayısı kullanılmıştır. Skor hesabında bu katsayının kullanılması ile yakın geçmiş etkisinin daha yüksek önem taşıması gerçekleşmiştir. Böylece herhangi bir düğüm çifti arasında zaman etkisi ve bağlantı ağırlığının da kullanıldığı bir hesaplama yapılmış olmaktadır.

Deneysel çalışmalarda α değeri 0,1 olarak belirlenmiştir. Çalışmada gözetimli yöntem önerilmiştir. Sınıflandırıcı olarak WEKA C4.5 karar ağacı algoritması yöntemi tercih edilmiştir. Önerilen modelin başarımını göstermek için yapılan deneysel çalışmada veri kümesi olarak 2007-2009 yılları arasındaki 60.000 kullanıcıdan oluşan facebook kullanıcı ağı ve iki farklı nitelik kümesi kullanılmıştır. Birinci nitelik kümesi oluşturulurken PropFlow ölçütü, ikinci küme ise önerilen T_Flow ve PropFlow ölçütleri ile oluşturulmuştur. Veri kümesi eğitim ve test kümesi olmak üzere ikiye ayrılmış ve C4.5 karar ağacı algoritması kullanılarak gözetimli öğrenme yöntemi ile önerilen yöntemin başarımı ölçülmüştür. Başarımının PropFlow'a göre daha yüksek olduğu F-ölçütü değeri ile gösterilmiştir. Yakın geçmiş

etkisinin kullanıldığı önerilen yöntemde PropFlow’a göre %4 oranında daha fazla yüksek başarımla sağladığı iddia edilmektedir.

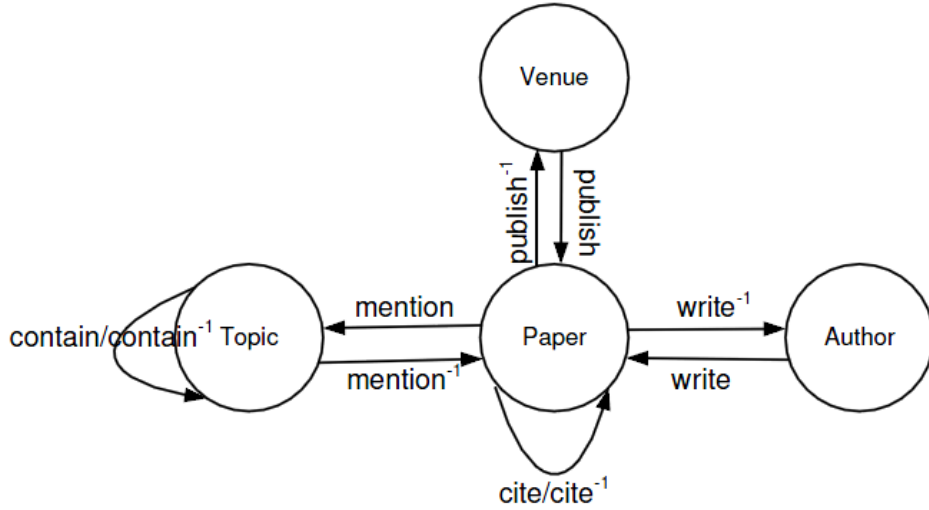
2.1.3 Heterojen ağlarda bağlantı tahmini

Ağ üzerindeki bağlantı tahmini üzerine birçok model önerilmiştir. Bu modeller, düğümler ve bu düğümler arası bağlantıların aynı türlerden oluştuğu homojen ağlar için veya düğümler ve/veya bağlantıların farklı türlerden oluştuğu heterojen ağlar için önerilmiştir. Sun ve diğ. (2011), Davis ve diğ. (2012), Johnson ve diğ. (2013), Lee ve diğ. (2012) çalışmaları ise heterojen ağlar için bağlantı tahmininde farklı çözümler önermiş ancak zaman bilgisinden yararlanmamışlardır. A düğüm tipleri kümesini, R ilişki tipleri kümesini göstermek üzere; heterojen ağlar $|A| > 1$ veya $|R| > 1$ koşuluna uyan ağlar olarak tanımlanmaktadır.

Çalışmada farklı tür düğümler ve farklı bağlantılar içeren heterojen ağlar için PathPredict (Sun ve diğ, 2011) yöntemi önerilmiştir. Önerilen yöntemde heterojen bir ağda verilen iki düğüm arası yakınlık değeri olası yolların anlamsal topolojik özellikleri (“meta path-based topological features”) ile hesaplanmaktadır. Bu doğrultuda düğüm çiftleri arasındaki yollarda kenar ağırlıklarının geçiş olasılıkları olarak kullanılması yerine, yolların anlamsal bilgisi kullanılarak yeni ağ tabanlı yakınlık önerilir. Yayın düğümleri arasındaki bağlantı tahmini için, yayınların yazarları arasındaki “ortak yayın yapma” bağlantıları tahmin edilmiştir. Bu çalışma için gerekli veri, DBLP bibliyografik ağ sitesinden toplanmakta ve 1989 ile 2009 yılları arasında bilgisayar bilimleri alanında yayınlanan makalelere ilişkin bilgileri içermektedir.

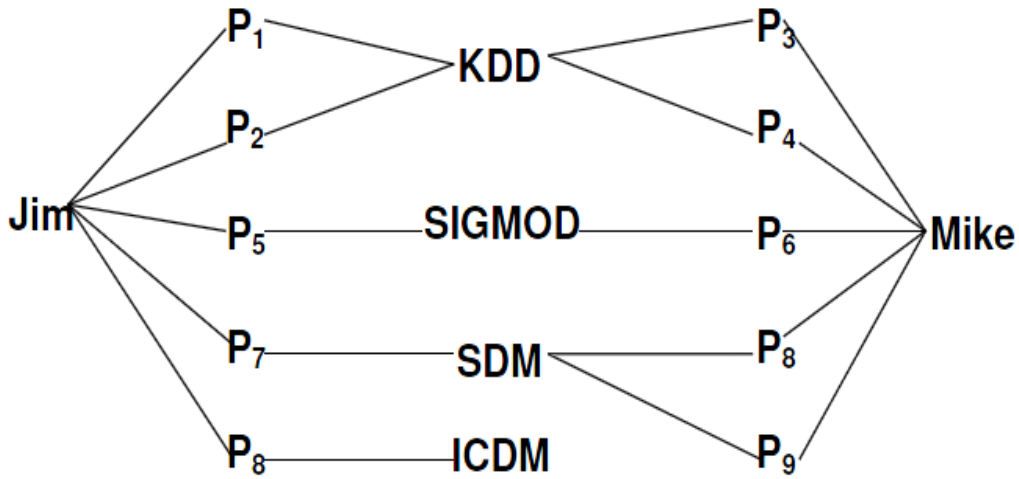
Çalışmada oluşturulan farklı tür düğümler ve farklı bağlantılar içeren heterojen ağ yazar (“author”), konferans yada dergi (“venue”), konu (“topic”) ve yayın (“paper”) olmak üzere dört farklı düğüm tipinden oluşmaktadır. Düğümler arasındaki bağlantı türleri ise yazarlar ve yayınlar arasında yazma (“write”), tarafından yazılma (“written by”); yayınlar arasında atıfta bulunma (“cite”), atıfta bulunulma (“cited by”); yayınlar ve konferanslar arasında yayınlanma (“publish”), tarafından yayınlanma (“published by”) v.b olmak üzere Şekil 2.3’teki ağ şemasında verilmektedir.

Oluşturulan ağ yönlü bağlantılardan oluştuğu için her ilişki türünün terside bağlantı tahmininde kullanılmaktadır.



Şekil 2.3 : Heterojen bibliyografik ağ şeması (Sun ve diğ, 2011).

İki yazar arasındaki aynı konferansta yayın yapma ilişkisini içeren anlamsal yol “A-P-V-P-A” Şekil 2.4’teki gibi verilebilir. İki yazar arasında bu anlamsal yolu içeren örnek bir yolda “Jim-P5-SIGMOD-P6-Mike” şeklinde verilebilir. İki yazar yazmış oldukları P5 ve P6 yayınlarını SIGMOD konferansında yayınladıkları için “A-P-V-P-A” anlamsal yolu ile birbirlerine bağlanmışlardır.



Şekil 2.4 : A-P-V-P-A anlamsal yolu (Sun ve diğ, 2011).

Ortak yayın yapma tahmini için yazar tipindeki düğümler arasında bütün anlamsal yollar belirli bir uzunluk değeri ile kısıtlı olarak belirlenmiştir. Örneğin uzunluğu 4 ile sınırlı yazar tipindeki düğümler arası anlamsal yollar ve ilişki bilgileri Çizelge 2.1’deki gibi verilebilir.

Çizelge 2.1 : Anlamsal yol – ilişki bilgileri.

Anlamsal yol	İlişki bilgisi
$A - P - A$	a_i ve a_j ortak yayın yapmıştır.
$A - P \rightarrow P - A$	a_i a_j 'ye atıf yapmıştır.
$A - P \leftarrow P - A$	a_i 'ye a_j tarafından atıf yapılmıştır.
$A - P - V - P - A$	a_i ve a_j aynı konferansta yayın yapmıştır.
$A - P - A - P - A$	a_i ve a_j aynı yazar ile ortak yayın yapmıştır.
$A - P - T - P - A$	a_i ve a_j aynı konuda yayın yapmıştır.
$A - P \rightarrow P \rightarrow P - A$	a_i a_j 'ye atıf yapan yayına atıf yapmıştır.
$A - P \leftarrow P \leftarrow P - A$	a_i 'ye a_j 'nin atıf yaptığı yayın tarafından atıf yapılmıştır.
$A - P \rightarrow P \leftarrow P - A$	a_i ve a_j aynı yayına atıf yapmıştır.
$A - P \leftarrow P \rightarrow P - A$	a_i ve a_j 'ye aynı yayın tarafından atıf yapılmıştır.

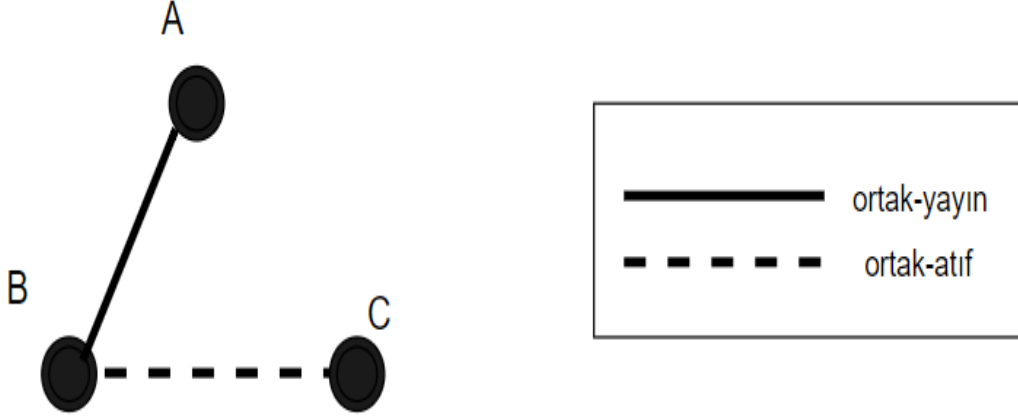
Çalışmada önerilen aynı tür düğümler arasında bulunan yolların yakınlığını temsil eden benzerlik metriği “*PathSim*” ise $p_{x \rightarrow y}$; x ve y düğümleri arasındaki bir yol örneğini, $p_{x \rightarrow x}$; x ve x düğümü ve $p_{y \rightarrow y}$; y ve y düğümleri arasındaki yol örneğini göstermek üzere denklem 2.31’e göre tanımlanmıştır.

$$sim(x, y) = \frac{2 * |\{p_{x \rightarrow y} : p_{x \rightarrow y} \in P\}|}{|\{p_{x \rightarrow x} : p_{x \rightarrow x} \in P\}| + |\{p_{y \rightarrow y} : p_{y \rightarrow y} \in P\}|} \quad (2.31)$$

Çalışmada gözetimli yöntem önerilmiştir ve bağlantıların (ortak yayın) olup olmadığı lojistik regresyon analizi ile belirlenmiştir. Düğüm çiftleri arasındaki yollara ait anlamsal topolojik özellikler sınıflandırıcıya girdi olarak verilmektedir. Sınıflandırıcı farklı anlamsal yolların (ilişki türleri) ağırlıklarını belirlemekte ve buna göre ortak yayın yapma tahmini yapılmaktadır. Yapılan deneysel çalışmalarda heterojen bibliyografik ağ’da anlamsal topolojik özelliklerinin kullanılmasının bağlantı (ortak yayın) tahmininde önemli iyileştirme sağladığı iddia edilmektedir.

Heterojen ağlarda yapılan yapılan bir diğer çalışmada da (Davis ve diğ, 2011) fenotipik ve genetik hastalıklar ağında bağlantı tahmini üzerine çalışma yapılmıştır.

Çalışmada farklı tür bağlantılar içeren heterojen ağlar için çok ilişkili bağlantı tahmini ölçütü önerilmiştir. Çalışmada gen hastalık etkileşim bilgileri kullanılarak heterojen ağ yapısı oluşturulmuştur (Davis ve diğ, 2013). Şekil 2.5’te verilen örnek heterojen ağ yapısında düğümler yazarları, düğümler arasındaki bağlantılar ise ortak-yayın ve ortak-atıf iki farklı türe ait ilişkileri ifade etmektedir.



Şekil 2.5 : Örnek heterojen ağ.

Yöntemde Adamic Adar ölçütü heterojen ağlar için genişletilmiş ve düğüm çiftleri arasında benzerliğin bulunması için önerilen skor hesabında farklı bağlantı türlerine ait düğüm derece bilgisi hesaplamada dikkate alınmıştır.

a ile c düğümleri arasında bağlantı türü x olan bağlantının skoru denklem 2.32'deki gibi bulunmaktadır. $derece_x(n)$ n düğümüne bağlı, türü x olan bağlantı sayısını diğer bir ifadeyle n düğümünün x bağlantı türüne göre derecesini göstermektedir.

$$skor_x(a, c) = \sum_{n \in N_a \cap N_c} \frac{1}{\log \left\{ \begin{array}{ll} derece_{type(a,n)} n & type(a,n)=type(c,n) \\ derece_{type(a,n)} n + degree_{type(c,n)} n & type(a,n) \neq type(c,n) \end{array} \right.} \quad (2.32)$$

Bu çalışma için gerekli olan gen ve hastalık verileri OMIM, Swiss-Prot, HPRD veritabanlarından toplanmakta ve veriler hastalık gen etkileşimi ve protein protein etkileşimine ilişkin bilgileri içermektedir. Testler sonucunda çalışmada heterojen ağlar için önerilen MRLP ölçütünün homojen ağlarda kullanılan statik ölçütlerden (AA, CN, JC, PA) daha iyi sonuç verdiği görülmüştür.

Fenotipik ve genetik hastalıklardan oluşan heterojen ağlarda yapılan bir diğer çalışmada da PropFlow ölçütünü heterojen ağlar için genişleten MRPF (Multi-Relational PropFlow) ölçütü önerilmiştir (Johnson ve diğ, 2013). Heterojen ağ yapısında düğümler hastalıkları, düğümler arasındaki bağlantılar ise fenotipik ve genetik iki farklı türe ait ilişkileri ifade etmektedir.

Farklı türlerdeki bağlantılar arasında korelasyonu ifade etmek için E_1 ve E_2 farklı bağlantı türlerini göstermek üzere $\sigma(E_1, E_2)$ katsayısı modele eklenmiştir. $\sigma(E_1, E_2)$ korelasyon katsayısı farklı bağlantı türlerinin birbirine olan etkisini tanımlamakta ve asimetrik özellik taşımaktadır; $\sigma(E_1, E_2) \neq \sigma(E_2, E_1)$ diğer bir ifadeyle bağlantı türü E_1 ve E_2 birbirlerini farklı derecelerde etkilemektedir. Örneğin insanların lokasyon bilgisi arkadaşlık bağlantısının bulunmasına katkıda bulunmakta fakat arkadaşlık bilgisi lokasyon bilgisinin bulunmasına aynı ölçüde katkı sağlamamaktadır. Çalışmada önerilen MRPF ölçütünde kenar ağırlıkları geçiş olasılıkları olarak kullanılmakta ve kaynak düğüm ile hedef düğüm arasında rassal gezinti olasılığı ile bilgi akışı değeri hesaplanmaktadır.

PropFlow ölçütünde olduğu gibi MRPF ölçütünde de ağda gezinti için belirli bir derinlik değeri ile kısıtlı enine öncelikli arama algoritması tercih edilmiştir. Ölçüt genel olarak belirli bir düğümden başlayıp bir başka düğümde biten belirli uzunluktaki kısıtlı bir rassal gezintinin olasılığı ile ilgilidir. s kaynak düğüm, t hedef düğüm, i bağlantı türü olmak üzere bilgi akışı değeri denklem 2.33'teki gibi hesaplanmaktadır. PropFlow'dan farklı olarak bilgi akış değeri her bir bağlantı türü için hesaplanmaktadır.

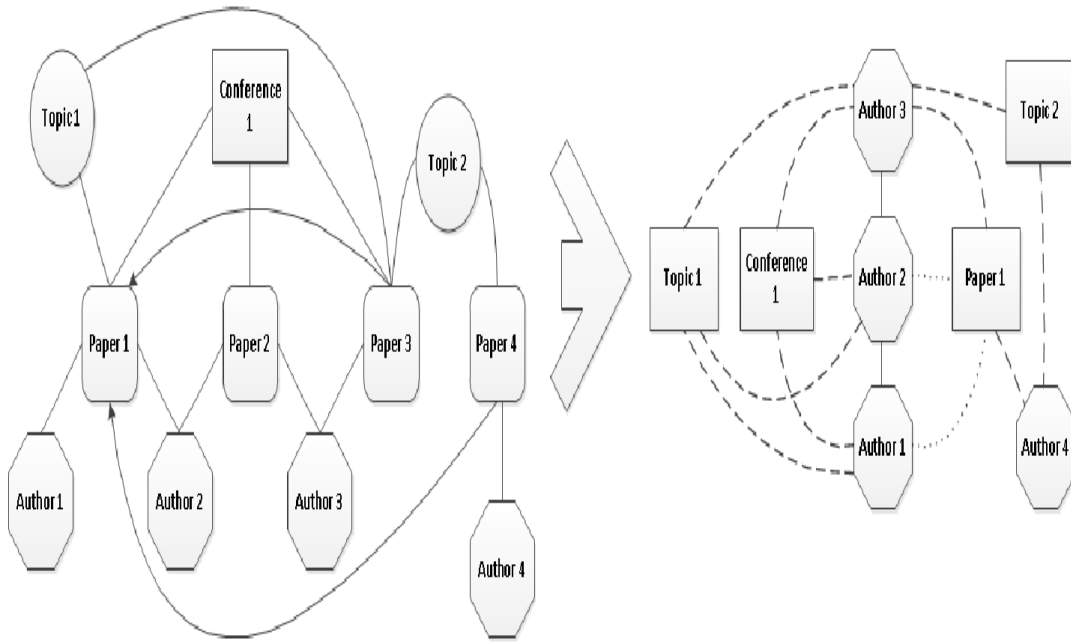
$$flow(s, t, i) = score(s) * \beta^{h-1} * \frac{w(s, t, i)}{d(s, i)} + \beta^{h-1} * p(i) * (1 - p(i)) * \frac{\sum_{k \neq i}^K p(i/k) * \frac{w(s, t, k)}{d(s, k)}}{K-1} \quad (2.33)$$

$w(s, t, i)$, s kaynak düğümü ile t hedef düğümü arasındaki bağlantı türü i olan kenarın ağırlığını, d kaynak düğümünün derecesini, K kaynak düğüm s 'e bağlı toplam bağlantı türü sayısını, $p(i/k)$ bağlantı türleri arasındaki korelasyon değerini, h kaynak ve hedef düğüm arasındaki en kısa adım sayısını göstermektedir. $score(s)$ değeri ise s düğümüne gelen olasılık değeridir ve başlangıç düğümü için değeri 1'dir. Modelde β değeri sabit 0,05 olarak alınmıştır.

Bu çalışma için gerekli olan gen ve hastalık verileri OMIM, Swiss-Prot, HPRD veritabanlarından toplanmakta ve veriler hastalık gen etkileşimi ve protein protein etkileşimine ilişkin bilgileri içermektedir. Testler sonucunda çalışmada heterojen ağlar için önerilen MRPF ölçütü homojen ağlarda kullanılan statik ölçütlerden (AA, CN, JC, PA) ve heterojen ağlar için önerilen MRLP ölçütünden daha iyi sonuç verdiği görülmüştür.

Heterojen ağlarla ilgili bir diğer çalışmada farklı tür düğümler ve farklı bağlantılar içeren heterojen ağlar için rassal gezinti tabanlı bağlantı tahmini (Lee ve Adorna, 2012) yöntemi önerilmiştir. Yayın yazarları arasındaki “ortak yayın yapma” bağlantılarının tahmin edilmesi gerçekleştirilmiştir. Önerilen yöntemde öncelikle oluşturulan farklı tür düğümler ve farklı bağlantılar içeren heterojen ağın yapısı modifiye edilmiştir. Yazar (“author”), konferans yada dergi (“venue”), konu (“topic”) ve yayın (“paper”) olmak üzere dört farklı düğüm tipinden oluşan heterojen ağ, yazar düğümü (“author node”) ve özellik düğümü (“attribute node”) iki farklı tipten oluşan ağ yapısına çevrilmiştir (Şekil 2.6).

Özellik düğümleri konferans ya da dergi (“venue”), konu (“topic”) ve atıf almış yayın (“cited paper”) türlerinden oluşmaktadır. Oluşturulan modifiye ağda yazar çiftleri yapısal bağlantılar (structural links), yazar düğümleri ile özellik düğümleri ise özellik bağlantıları (attribute links) ile bağlanmıştır. Özellik bağlantıları yazarların aynı konferansta yayın yapma, aynı konuda yazma ve aynı yayına atıfta bulunma özelliklerini taşıdığı için ortak yayın yapma tahminine katkı sağlamaktadır.



Şekil 2.6 : Örnek heterojen ve modifiye ağ (Lee ve diğ, 2012).

Yöntemde bağlantı ağırlıkları geçiş olasılıkları olarak kullanılmakta ve kaynak düğüm ile hedef düğüm arasında rassal gezinti olasılığı ile benzerlik değeri hesaplanmaktadır. Bağlantıları ağırlıklandırmak için etki ölçütleri (“importance measures”) tanımlanmıştır ardından tanımlanan bu ölçütlere göre yapısal bağlantılar ve özellik bağlantıları ağırlıklandırılmıştır. Yazarlar arasındaki yapısal bağlantılar ve özellik bağlantıları denklem 2.34 ve denklem 2.35’teki eşitlikler kullanılarak ağırlıklandırılmıştır. Her bir etki ölçütünün ağırlık katsayısı α , β , γ , δ değerleri deneysel olarak belirlenmiştir. $w_a(x)$ yazar ve özellik düğümleri arasındaki özellik bağlantısının ağırlığını, $w_a(a')$ ise yazarlar arasındaki yapısal bağlantının ağırlığını göstermektedir.

$$w_a(x) = \alpha g(x) + \beta l_a(x) + \gamma f_a(x) + \delta r_a(x) \quad (2.34)$$

$$w_a(a') = \left(\frac{\alpha}{3} + \beta\right) l_a(a') + \left(\frac{\alpha}{3} + \gamma\right) f_a(a') + \left(\frac{\alpha}{3} + \delta\right) r_a(a') \quad (2.35)$$

Bağlantıların ağırlıklandırılmasında kullanılan etki ölçütleri ise şu şekilde tanımlanmıştır.

Genel etki ölçütü, herhangi bir x özellik düğümüne bağlı ortak yayın yapan yazar çiftlerinin, muhtemel ortak yayın yapan yazar çifti sayısına oranı şeklinde tanımlanmaktadır. Denklem 2.36’da n_x , x özellik düğümünün derece sayısını göstermektedir. a ve a' yazarları ortak yayın yapmış ve x düğümüne bağlı ise $e_x(a, a')$ 1 değerini almaktadır.

$$g(x) = \frac{\sum_{(a,a') \in E_I} e_x(a, a')}{\binom{n_x}{2}} \quad (2.36)$$

Yerel etki ölçütü, herhangi bir x özellik düğümünün a yazar düğümünün ortak yayın yaptığı yazarlara göre önemini göstermektedir. $N_{auth}(a)$, a yazarının ortak yayın yaptığı yazarları göstermektedir (denklem 2.37). a yazarının ortak yayın yaptığı yazar düğümleri x düğümüne bağlı ise $A(a', x) = 1$ değerini almaktadır.

$$l_a(x) = \sum_{a' \in N_{auth}(a)} A(a', x) \quad (2.37)$$

Sıklık etkisi ölçütü, $freq(a, x)$ değeri a ve x düğümlerinin bağlantı sayısı olmak üzere denklem 2.38’e göre hesaplanmaktadır. k değeri 1,1 olarak alınmıştır.

$$f_a(x) = k^{freq(a, x)} \quad (2.38)$$

Yakın geçmiş etkisi ölçütü; $rec(a,x)$ değeri a ve x düğümlerinin en son etkileşim zamanını göstermek üzere denklem 2.39'a göre hesaplanmaktadır. l_y ağdaki en son bağlantının kurulma zamanıdır. Örneğin a düğümü yazar ve x düğümü bildiri ise l_y değeri ağdaki en son yayınlanan bildirinin tarihini göstermektedir bu durumda $l_y - rec(a,x)$ değeri azaldıkça yakın geçmiş etkisi büyük değerler almaktadır.

$$r_a(x) = \frac{1}{k^{l_y - rec(a,x)}} \quad (2.39)$$

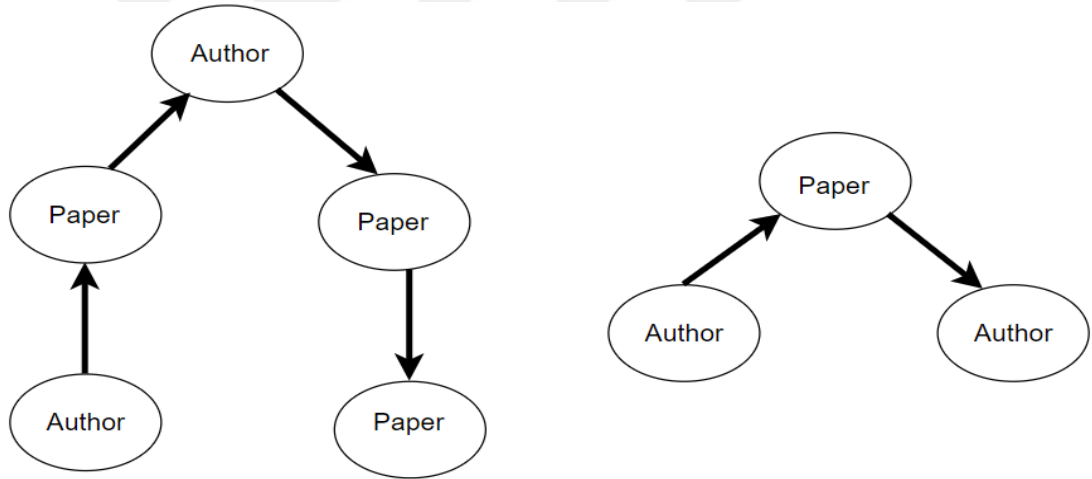
Çalışma için gerekli veri, DBLP bibliyografik ağ sitesinden toplanmakta ve 1989 ile 2009 yılları arasında bilgisayar bilimleri alanında yayınlanan makalelere ilişkin bilgileri içermektedir. Ağ bağlantıları ağırlıklandırıldıktan sonra rassal yürüyüş algoritması ile rassal gezintinin olasılık değeri hesaplanmakta ve yazarlar arasındaki “birlikte yayın yapma” ilişkileri tahmin edilmektedir.

Bir diğer çalışmada (He ve diğ., 2014) farklı tür düğümler ve farklı bağlantılar içeren heterojen ağlar için PathPredict yöntemi geçişlilik (transivity) özelliği ile genişletilmiştir. En genel hali ile geçişlilik özelliği ağlardaki üçlü (triad) yapılar bağlı olmak üzere herhangi bir i ve j düğümü ve j ve h bağlı ise i ve h düğümleri arasında da bağlantı oluşma durumu şeklinde tanımlanmaktadır (arkadaşımın arkadaşı arkadaşımıdır ilkesi). Önerilen yöntemde heterojen bir ağda verilen iki düğüm arası yakınlık değeri olası yolların anlamsal topolojik özellikleri (meta path-based topological features) ile hesaplanmaktadır. Bu doğrultuda düğüm çiftleri arasındaki yollarda kenar ağırlıklarının geçiş olasılıkları olarak kullanılması yerine, yolların anlamsal bilgisi kullanılarak yeni ağ tabanlı yakınlık önerilir. Yayın yazarları arasındaki “ortak yayın yapma” ilişkilerinin tahmin edilmesi gerçekleştirilmiştir. Bu çalışma için gerekli veri, DBLP bibliyografik ağ sitesinden toplanmakta ve 1989 ile 2009 yılları arasında bilgisayar bilimleri alanında yayınlanan makalelere ilişkin bilgileri içermektedir.

Çalışmada oluşturulan farklı tür düğümler ve farklı bağlantılar içeren heterojen ağ yazar, konferans ya da dergi, konu ve yayın olmak üzere dört farklı düğüm tipinden oluşmaktadır. İki yazar arasındaki aynı yazar ile yayın yapma ilişkisini içeren anlamsal yol “yazar-yayın-yazar-yayın-yazar (A-P-A-P-A)” şeklinde ve ortak yayın anlamsal yolu ise “yazar-yayın-yazar (A-P-A)” şeklinde Şekil 2.7'deki gibi verilebilir.

Çalışmanın temel düşüncesi Sun ve diğ. (2011) önerdiği *Pathsim* benzerlik metriğinin geçişlilik özelliği ile genişletilmesidir bunun için anlamsal yollar alt anlamsal yollara bölünerek geçişlilik özelliği ile yolların ağırlıklandırılması amaçlanmıştır.

Diğer bir ifade ile T_1 ve T_l aynı türlerde düğümler olmak üzere anlamsal yol $P = T_1 \dots T_{m1} \dots T_{m2} \dots T_{md} \dots T_l$ şeklinde tanımlanmaktadır. Önerilen yöntem ile T_1 ve T_l düğümleri arasındaki P anlamsal yolu $P_1 = T_1 \dots T_{m1}$, $P_2 = T_{m1} \dots T_{m2}$, ..., $P_d = T_{md-1} \dots T_{md}$, $P_{d+1} = T_{md} \dots T_l$ olmak üzere $d+1$ adet alt yola bölünmekte ve her alt yolunda başlangıç ve bitiş düğümleride aynı tür düğümlerden oluşmaktadır. Alt yollar düğümlerin benzerliklerine göre ağırlıklandırılarak T_1 ve T_l düğümü arasındaki geçişli benzerlik (transitive similarity) ağırlıklı alt yollar ile hesaplanmaktadır. Önerilen yöntemde anlamsal yolları alt anlamsal yollara bölerek geçişlilik özelliğinin eklenmesi PathSim'e göre daha iyi sonuçlar vermiştir.



Şekil 2.7 : APAPA ve APA anlamsal yolu (He ve diğ, 2014).

Cao ve diğ. (2014) kapsamında, düğümler arasındaki benzerliği ölçmek için ilişkililik ölçütü (relatedness measure) önerilmiş ve bu ölçüt kullanılarak farklı bağlantı türleri kolektif olarak tahmin edilmeye çalışılmıştır. Önerilen bu ölçüt ile düğümler arasındaki yolların anlamsal topolojik özellikleri genişletilerek farklı tür düğümler arasındaki ilişkiler benzerlik hesabına katılmıştır. Önerilen model gen, doku, hastalık ve kimyasal bileşen tipinden düğüm tipleri ile aralarındaki farklı ilişkilerden oluşan heterojen ağda test edilmiş ve bağlantı tahmini başarımının önerilen kolektif öngörü yöntemi ile artırıldığı görülmüştür.



3. MATERYAL VE METOT

3.1 Karmaşık Ağlar ile İlgili Genel Kavramlar

Karmaşık Ağ modellerinin özellikleri incelendiğinde karmaşık ağlar ile ilgili yapılan çıkarımlarda küçük dünya hipotezi, demetleme katsayısı ve derece dağılımı olmak üzere üç kavramın önemli rol oynadığı ifade edilmektedir (Albert ve Barabasi, 2002).

3.1.1 Küçük dünya hipotezi

Küçük dünya ağları; ağın kendi boyutuna oranla, herhangi iki düğüm arasındaki en küçük yol uzunluğunun kısa ve demetleme sayısının beklenenden büyük olduğu ağlardır. En kısa yolların en uzun olan yol ağın çapı şeklinde tanımlanmakta ve ağı karakterize eden faktörlerden biri olarak değerlendirilmektedir. Buna göre küçük dünya ağları küçük çap değerine sahiptir. Küçük dünya ağlarında düğümlerin dereceleri ağın ortalama derecesine göre fazla sapma göstermemekte ve derecelerin varyansı küçük değere sahip olmaktadır. Küçük dünya hipotezine, Paul Erdős'e makale yazma uzaklık değeri olarak tanımlanan Erdős sayısı örnek olarak gösterilebilir. Paul Erdős ile makale yazarların Erdős Sayısı 1 iken, Erdős ile makale yazmış yazar ile makale yazarların sayısı 2'dir. 268.000 yazardan oluşan ağın ortalama Erdős sayısı 4,65 olarak belirlenmiştir. Benzer şekilde küçük dünya ile ilgili en popüler gösterim altı derecelik ayırım kavramı ile Milgram (1967) tarafından ortaya konulmuştur. Buna göre Amerika'da insanlar arasındaki tanışma ilişkisini temel alan sosyal ağda, herhangi bir düğüm çifti arasındaki en kısa ortalama uzaklık 6 olarak bulunmuştur.

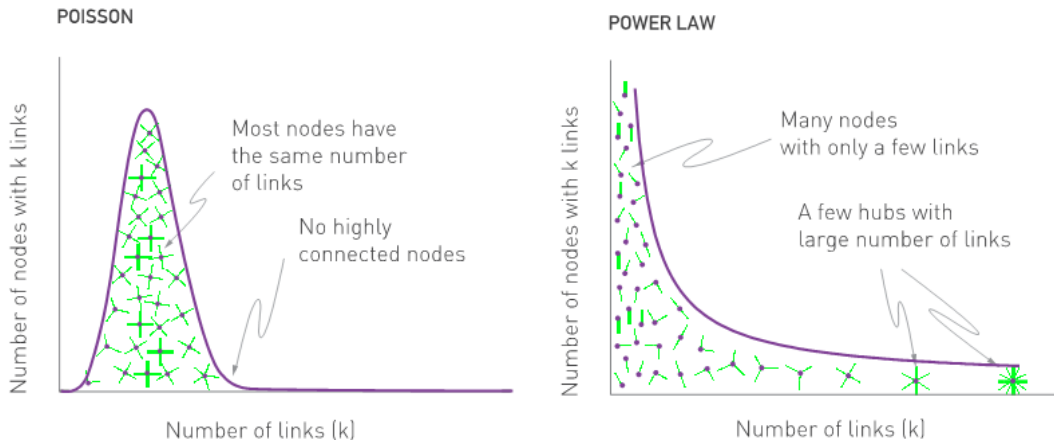
3.1.2 Demetleme katsayısı

Watts ve Strogatz (1998) demetlemeyi ölçmek için kullanılan demetleme katsayısı olarak adlandırılan ölçüyü geliştirmişlerdir. Demetleme katsayısı ağda yer alan her bir düğüm için ayrı ayrı hesaplanabileceği gibi ağın bütünü içinde hesaplanabilmektedir. Bütün düğümlerin birbirine bağlı olduğu tam bağlantılı ağda demetleme katsayısı 1 değerini almaktadır. Rassal bir ağda ise bağlantılar rassal dağıldığı sürece demetleme katsayısı bağlantı olasılığı p değeri ile aynı değeri

almaktadır. Derecesi k_i olan bir i düğümü için lokal demetleme katsayısı, L_i değeri k_i komşu arasındaki gerçekleşmiş bağlantıları göstermek üzere $C_i = \frac{2L_i}{k_i(k_i-1)}$ şeklinde ifade edilmektedir. Buna göre ağın ortalama demetleme katsayısı $\frac{1}{N} \sum_{i=1}^N C_i$ şeklinde belirlenir. Watts ve Strogatz demetleme katsayısı ölçüsünü ve ortalama en kısa yol uzunluğu ölçüsünü ağların sınıflandırılması için kullanmışlardır. Küçük dünya ağları aynı düğüm ve kenar sayısına sahip diğer rassal ağlarla karşılaştırıldığında demetleme katsayısının beklenenden daha büyük ve ortalama en kısa yol uzunluğunun ise beklenenden daha kısa olduğu görülmektedir.

3.1.3 Derece dağılımı

Bir ağın bütün düğüm derecelerinin dağılımı derece dağılımı olarak adlandırılır ve ağın bütününe ilişkin bir özelliği göstermektedir. Derece dağılımı $p(k)$, N_k derecesine sahip düğümleri ve N toplam düğüm sayısını temsil etmek üzere $p(k) = \frac{N_k}{N}$ şeklinde ifade edilmektedir. Ayrıca derece dağılımı ile ağın nasıl organize edildiği açıklanmaya çalışılmaktadır. Örneğin Poisson dağılımı düğümlerin rassal ve eşit olasılıklı bağlantılar ile bağlandığını gösterirken, power-law derece dağılımı düğümlerin tercihli olarak bağlandığını diğer bir ifade ile yeni düğümlerin çok sayıda bağlantıya sahip düğümlere bağlanma olasılıklarının yüksek olduğunu gösterir (Şekil 3.1).



Şekil 3.1 : Poisson ve Power Law derece dağılımları (Barabási, 2016).

3.1.4 Rassal çizge modeli

Bir rassal ağ n adet düğümden oluşan bir düğüm kümesi ile başlar ve bu düğümlere rassal olarak kenarların eklenmesi ile elde edilir. Erdős ve Renyi (1959) rassal çizgeyi $\frac{n(n-1)}{2}$ adet mümkün bağlantıdan rassal olarak seçilen N adet bağlantının n adet düğümü birbirine bağlaması olarak tanımlamakta ve stokastik bir süreç olarak ifade etmektedir (Erdős ve Renyi, 1959). Rassal Çizge Modeli'ne göre her bir düğümün başka bir düğümle bağlantı kurma olasılığı eşittir ve bu bağlantılar rassal bir süreç şeklinde meydana gelmektedir. Bu süreç n adet düğüm ile başlar ve rassal çizgeye rassal bağlantıların eklenmesi ile gelişir. Sürecin sonunda maksimum bağlantı sayısına $\frac{n(n-1)}{2}$ ulaşılır ve tam bağlı bir ağ elde edilir. Rassal ağlarda düğümlerin dereceleri p bağlanma olasılığı olmak üzere, $p(n-1)$ ortalama etrafında normal dağılır. Ancak gerçek dünya ağlarının bu şekilde olmadığı Barabasi ve Albert (1999)'in çalışmaları ile anlaşılmış ve gerçek dünya ağlarında ölçekten bağımsız ağ örneklerine rastlanmıştır.

3.1.5 Ölçekten bağımsız model

Barabasi ve Albert'in (1999) küçük dünya deneyi ile ilgili web'te yaptığı çalışmalar ile sayfalara gelen ve sayfalardan çıkan bağlantı sayılarının incelenmesi ile web'in yapısında yüksek dereceli birkaç düğümün egemen olduğu ve web sayfalarındaki bağlantıların dağılımının bir kuvvet yasası (power-law) izlediği belirlenmiştir. Daha sonra literatürde birçok ağın bu yasaya uyduğu bulunmuş ve power-law dağılımını izleyen ağlar ölçekten bağımsız ağlar olarak adlandırılmıştır.

Power-law dağılım eğrisi rassal ağ modelinin sahip olduğu Poisson dağılım eğrisi gibi bir tepe noktasına sahip değildir. Poisson dağılım eğrisine sahip rassal ağ modelini temsil eden bir ağda, düğümlerin çoğunluğu aynı sayıda kenara sahiptir. Bu nedenle, rassal ağın düğümlerinin bağlanabilirliğinde karakteristik bir ölçek vardır. Bu karakteristik ölçek, ortalama düğüm sayısı tarafından şekillendirilmekte ve derece dağılımının tepe noktası tarafından belirlenmektedir. Power-law dağılım eğrisine sahip bir ağda ise, çok sayıda düğüm az sayıda bağlantıya, az sayıda düğüm ise çok sayıda bağlantıya sahip olmaktadır. Bu şekilde dağılım eğrisinde Poisson dağılım eğrisindeki gibi bir tepe noktası yoktur. Power-law dağılımının tepe noktasının olmaması, gerçek bir ağda karakteristik bir düğümün olmadığı anlamına gelmektedir.

diğer bir ifade ile bütün düğümler için karakteristik olduğu öne sürülebilecek bir düğüm yoktur (Barabasi ve Albert, 1999). Bu nedenle Barabasi – Albert modeli ölçekten bağımsız model olarak adlandırılmıştır. Büyük ağların derece dağılımı $P(k) \sim k^{-\gamma}$ ile ifade edilen power-law dağılımı göstermektedir (Gürsakal, 2009).

Barabasi ve Albert power-law dağılımının ortaya çıkışını tercihli bağlanma olarak adlandırdıkları mekanizma ile açıklamaktadır. Ağın büyümesi ile her seferinde ağı katılan yeni bir düğümün bağlanacağı düğüme karar vermesi durumunu ifade eder. Her yeni düğümün daha fazla bağlantıya sahip düğümü tercih etmesi ile ağın ilk düğümleri çok yüksek sayıda bağlantı sahibi olur. Böylece tercihli bağlanma, zengin daha zenginleşir olgusunu harekete geçirir ve ilk düğümlerin daha sonra gelenler aleyhine yüksek dereceye sahip olmasına neden olur.

3.2 Zaman Serileri

Zaman serileri; istatistik, ekonometri, finans gibi birçok alanda kullanılmaktadır. Serinin gelecek değerleri hakkında tahminler yapmak için, zaman içindeki yapısını belirlemek ve geçmişte gösterdiği davranışı ortaya çıkarmak gereklidir. Geçmişte ortaya çıkan bilgi yardımıyla, zaman serisini oluşturan değişken veya değişkenlerin gösterdiği davranışın belirlenmesi, yapılacak tahminlerin başarısını beraberinde getirmektedir (Pindyck ve Rubinfeld, 1991).

Eğer bir zaman serisi analizi tek değişkenin zaman içindeki hareketini inceliyorsa tek değişkenli zaman serisi, eğer birden fazla değişkenin birlikte zaman içindeki değişimini gözlemliyorsa çok değişkenli zaman serisi adını alır. Zaman serisi analizinin önşart olarak ele aldığı varsayımlarından en önemlisi serinin durağan (stationary) olmasıdır. Bir zaman serisinin durağan olması, zaman içinde belirli bir değere doğru yaklaşması veya beklenen değeri etrafında dalgalandığını ifade etmektedir. Durağanlık tüm gecikmeler için ortalama, varyans ve kovaryans değerlerinin sabit olmasıdır. Diğer bir ifadeyle serinin, sabit bir ortalama, sabit bir varyans ve gecikme seviyesine bağlı kovaryansa sahip olması şeklinde tanımlanmaktadır. Bir değişkenin durağan bir süreç izleyip izlememesinin önemi ise şu şekilde ifade edilebilir. Eğer bir stokastik süreç durağan değilse, serinin davranışı sadece ele alınan tahmin dönemi için geçerli olacaktır. Ancak seri hakkında diğer dönemler için bir genelleme yapılamayacaktır (Akgül, 2003).

Zaman serilerinde zaman içindeki değişme eğiliminin trend, mevsimsellik, konjonktürel ve düzensiz dalgalanmalar anlamında ayrıştırılması önemlidir. $Y = T + C + S + I$ olarak ifade edilen modelde, T; trend, C; konjonktürel hareketler, S; mevsimsellik olmak üzere modelin deterministik kısmını, I ise stokastik kısmını ifade etmektedir. Bir zaman serisi dört bileşenden birini veya birkaçını barındırabilir (Bozkurt, 2013).

3.2.1 VAR çok değişkenli zaman serisi modeli

VAR modeli, seçilen bütün değişkenleri birlikte ele alır, doğrusal bir modeldir ve bir sistem bütünlüğü içinde değişkenleri inceler. Değişkenler arası ilişkiler hakkında bir ön kısıt konulmaz.

İki değişkenli VAR modeli, standard şekilde denklem 3.1 ve 3.2’de belirtildiği gibi ifade edilebilir;

$$y_t = a_1 + \sum_{i=1}^p b_{1i}y_{t-i} + \sum_{i=1}^p b_{2i}x_{t-i} + \varepsilon_{1t} \quad (3.1)$$

$$x_t = c_1 + \sum_{i=1}^p d_{1i}y_{t-i} + \sum_{i=1}^p d_{2i}x_{t-i} + \varepsilon_{2t} \quad (3.2)$$

Modelde p gecikmelerin uzunluğunu, ε ortalaması 0, kendi gecikmeli değerleriyle olan kovaryansları 0 ve varyansları sabit, normal dağılıma sahip, rassal hata terimlerini göstermektedir.

Evrimsel heterojen ağların kompleks yapısı içinde farklı türlere ait bağlantılar arası ilişkileri açıklamak ve ağın topolojisini yorumlamak zordur. Ağı oluşturan düğümler ve bağlantılar karşılıklı olarak sürekli bir etkileşim halindedir. Bu nedenle herhangi bir topolojik değişkeni (benzerlik ölçütünü) tek denklemlerle bir modelde bağımlı değişken olarak belirleyerek modellemek yeterli olmayabilir. Çünkü kompleks yapı nedeniyle, ele alınan bağımlı değişken, bağımsız olarak düşünülen değişken ya da değişkenlerin belirleyicisi olabilir. Bu nedenlerle, sözü edilen bu karmaşıklık, değişkenlerin tek denklemlerle modeller yerine eşanlı denklem sistemleri ile incelenmesi gereğini ortaya koymaktadır. VAR modeli değişkenler arasında dinamik ilişkileri verebilen ve böylelikle zaman serileri alanında değişkenler arasındaki ilişkileri ve geleceğe yönelik tahminleri açıklayabilen bir yöntemdir (Kilian ve Chang, 2000).

VAR modeli, sistemde yer alan çok sayıda değişkenin geçmiş değerleri ile ifade edildiği ve her bir denklemin EKK (En Küçük Kareler) ile çözümlendiği bir çözüm

teknikidir. Geçmiş bilginin kullanılması ile sağlanan bu yapı zaman serilerine dinamiklik katarak, ilişkileri yorumlama konusunda bir açılım yaratmıştır. VAR modelleri ile tahmin yapabilmek için, sistemde yer alan tüm değişkenlerin durağan olmaları gerekmektedir (Lutkepohl, 1982).

Örnek olarak iki değişkenin yer aldığı, birinci mertebeden VAR(1) modeli denklem 3.3 ve 3.4’te belirtildiği gibi tanımlanabilir.

$$y_t = a_1 + b_{11}y_{t-1} + b_{21}x_{t-1} + \varepsilon_{1t} \quad (3.3)$$

$$x_t = c_1 + d_{11}y_{t-1} + d_{21}x_{t-1} + \varepsilon_{2t} \quad (3.4)$$

- i. y_t ve x_t durağan serilerdir.
- ii. ε_{1t} ve ε_{2t} sırasıyla σ_y ve σ_x standart sapmalı beyaz gürültü süreçlerdir.
- iii. ε_{1t} ve ε_{2t} kendi aralarında korelasyonsuzdur.

Modelde, y ve x değişkenleri durağan, hata terimleri ise beyaz gürültü sürecine sahiptir. Dikkat edilirse, VAR modelleri sistemdeki her değişkenin hem kendi gecikmeli değerleri, hem de sistemdeki diğer tüm değişkenlerin gecikmeli değerlerinin doğrusal bir fonksiyonudur. Bu durumda herhangi bir kısıtlama olmazsa VAR modelindeki her eşitliğin sağ tarafı aynı değişkenleri içerecektir.

Örneğin, düğümlerinin yazarları temsil ettiği ve “ortak yayın yapma” ve “ortak yayınlara atıf yapma” ilişkilerinin bulunduğu heterojen bir ağ olsun. Bu ağda herhangi bir i,j düğüm çifti için y_t ve x_t sırasıyla; t zamanındaki ortak yayın yapma ve ortak yayınlara atıf yapma ilişkilerine ait bağlantı sayılarıdır. Bu durumda i,j düğümleri arasındaki ortak yayın sayısı önceki dönemin hem ortak yayın sayısına hem de ortak yayın atıf sayısına bağlıdır. Ayrıca iki değişken arasında geri besleme ilişkisi vardır ve dolayısıyla ortak atıf sayısı da bir dönem önceki ortak yayın sayısından etkilenecektir.

Düğümler arasındaki etkileşimin karmaşıklığı ve çok yönlülüğü, eşanlı denklem sistemlerinin kullanılmasını zorunlu hale getirmektedir. Karmaşık ağlarda düğümler ve bağlantılar arasındaki karşılıklı etkileşimin doğal bir sonucu olarak ortaya çıkan bağımlı ve bağımsız değişkenin belirlenmesi aşamasında karşılaşılan güçlükler, analizin tutarlılığını da önemli ölçüde etkilemektedir. Eşanlı denklem sistemlerinin içerdiği bu karmaşık tablonun çözümüne yönelik olarak geliştirilen VAR modelleri ile söz konusu sorun ortadan kaldırılmaktadır. VAR modelleri, yapısal modele

herhangi bir kısıtlama getirmeksizin dinamik ilişkileri verilebildiği için zaman serileri açısından sıklıkla tercih edilmektedir (Akdi, 2012).

Çok değişkenli zaman serisi analizinde optimal gecikme uzunluğunun belirlenmesi için sıklıkla kullanılan kriterlerden biri AIC'dir. AIC yöntemi parametre sayısını azaltmakta ve zaman serileri ile tahmin yapılmasında yüksek başarımlı elde edilmesini sağlamaktadır. $|\sum_p m|$; p gecikme için hata kovaryans matrisinin determinantını, N gözlem sayısını, p gecikme uzunluğunu göstermek üzere AIC değeri denklem 3.5'te verilmiştir:

$$AIC(p) = \ln|\sum_p m| + \frac{2}{N} \quad (3.5)$$

AIC değerinin minimum kılın p değeri, denklem 3.5'te uygun gecikme seviyesi olarak belirlenir (Akaike, 1973).

3.2.2 AR, MA, ARMA / ARIMA tek değişkenli zaman serisi modelleri

Box-Jenkins yöntemi zaman serileri alanında sık kullanılan bir yöntemdir. 1970'li yıllarda temelleri atılan Box-Jenkins yöntemi durağanlığı, deterministik bileşen bilgisini ve geleceğe ilişkin tahminleri bir arada ortaya koyduğu için tercih edilen bir yöntemdir. Bir değişkene ilişkin yapılacak tahmin, kendi gecikmeli değerleri, hata terimleri ya da her ikisinin kombinasyonu ile yapılmaktadır (Box ve Jenkins, 1976).

Bu yöntemde üç modelleme söz konusudur. Bunlar AR Süreci, MA Süreci ve ikisinin birleşiminden oluşan Otoregresif Hareketli Ortalama (ARMA) Süreci'dir. Durağan olmayan bir seri için fark alınması gerektiğinde ise ARMA süreci Bütünleşik Otoregresif Hareketli Ortalama (ARIMA) süreci haline dönüşür. ARIMA modeli, farkı alınarak durağan hale getirilmiş otoregresif hareketli ortalama sürecini ifade etmektedir.

AR modelinde, denklem 3.6'da verilen eşitliğin sağ tarafı bağımlı değişkenin gecikmeli değerlerinden oluşur. Hata terimi ise rassal sürece sahiptir ve 0 ortalama ve sabit varyans değerine sahiptir.

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_p y_{t-p} + \varepsilon_t \quad (3.6)$$

Denklem 3.6'da eşitlik sabit parametrenin olmadığı ve bağımlı değişkenin değerinin sadece geçmiş değerleri ve hata terimi ile açıklandığı modeldir. Bu sürece AR(p)

süreci denilmektedir. p 'nin yani gecikmenin ne olacağına ilişkin karar minimum AIC değeri ile belirlenir.

Basit bir AR(1) süreci denklem 3.7'de ifade edildiği gibi ifade edilebilir.

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t \quad (3.7)$$

Bağımlı değişken, bir dönem gecikmesi, sabit parametre ve stokastik sürece sahip hata teriminden oluşmaktadır.

MA süreci, denklem 3.8'de ifade edildiği gibi bağımlı değişkenin, beyaz gürültü sürecine sahip geçmiş hata terimleri ile ifade edilmesidir.

$$y_t = \beta_0 + \beta_1 \varepsilon_t + \beta_2 \varepsilon_{t-1} + \beta_3 \varepsilon_{t-2} + \dots + \beta_q \varepsilon_{t-q} \quad (3.8)$$

Bu eşitlikte eşitliğin sağ tarafı anlamlı q sayısı kadar hata terimi ile ifade edilmektedir. Hata terimi, 0 ortalamaya ve sabit varyansa sahiptir.

ARIMA sürecinde (denklem 3.9), serilerin durağan bir yapı sergilemediği durumlarda seri ile ilgili geleceğe yönelik tahmin yapabilmek için, öncelikle serinin farkının alınması gerekir. Seri için durağan bir yapı elde edene kadar fark alma işlemi devam edecektir. d kadar farkı alınarak durağanlaşan seriye, d . mertebeden durağan seri adı verilir. Serinin durağanlaşması için fark alınması gerektiğinde (p,d,q) 'nın belirleneceği ARIMA yapısı geçerli olacaktır (Pindyck ve Rubinfeld, 1991).

$$y_t = \alpha_0 + \sum_{i=1}^p \alpha_i y_{t-i} + \sum_{i=0}^q \beta_i \varepsilon_{t-i} \quad (3.9)$$

Örneğin bir zaman serisini durağanlaştırmak için birinci ya da ikinci farkının alınması gerekmesi durumunda seriye ilişkin fark değerlerinin nasıl elde edildiği aşağıda özetlenmiştir.

Zaman serisinin birinci farkları,

$$\Delta x_t = x_t - x_{t-1}$$

Zaman serisinin ikinci farkları,

$$\Delta^2 x_t = \Delta x_t - \Delta x_{t-1}$$

Durağanlaştırma, d kare farkı alındıktan sonra ortalaması ve varyansı sabit, doğrusal bir otoregresif hareketli ortalama (ARMA) süreci sergileyen seri şeklinde tanımlanmaktadır (Engle ve Granger, 1987).

3.3 NARX Sinir Ağı Modeli

NARX (Nonlinear Autoregressive Exogenous – Doğrusal Olmayan Otoregresif Dışsal) modeline giriş olarak, tahmin edilmesi istenen değişkene (hedef değişken) ait geçmiş ölçüm verilerini içeren zaman serisi ve tahmin edilmesi istenen hedef değişkeni etkilemesi beklenen dışsal değişkenlere ait zaman serisi verilerek bir sonraki zaman dilimindeki hedef değişkenin değerinin tahmin edilmesi hedeflenmektedir. Bir $y(t)$ zaman serisinin, gelecekte alabileceği muhtemel değerler, $y(t)$ serisinin geçmişteki değerleri ve bunu etkileyen $u(t)$ dışsal değişkenler göz önüne alınarak, tahmin edilmeye çalışılmaktadır (Deljac ve diğ, 2016).

NARX modeli giriş, gizli ve çıkış katmanlarından oluşan ileri beslemeli, gecikme ve döngü mekanizmalarına sahip dinamik ağlardır. Bu nedenle zaman serisi analizlerinde sıklıkla kullanılmaktadır. NARX modeli ayrıca ARIMA, VAR zaman serisi modelleri gibi serinin durağan olmasını ve serinin doğrusal bir süreçten oluştuğunu önşart olarak kabul etmez bu nedenle doğrusal olmayan ve durağan olmayan zaman serilerini de modelleyebilmektedir (Junior ve diğ, 2008). NARX tabanlı sinir ağlarında, öğrenme algoritması olarak hız ve kararlılık sağlayan Levenberg-Marquardt (LMA) algoritması kullanılmaktadır. NARX modeline ait matematiksel gösterim denklem 3.10 ile ifade edilmekte ve hedef değişken $y(t)$, daha önceki değerleri ve dışsal giriş değişkeni $u(t)$ ile regresyona tabi tutulmaktadır (Liu ve diğ, 2013). $d_u \geq 1, d_y \geq 1, d_y \geq d_u$ olmak üzere, d_y hedef değişken için gecikme değerini, d_u dışsal değişken için gecikme değerini göstermektedir.

$$\hat{y}(t) = f(y(t-1), y(t-2), \dots, y(t-d_y); u(t-1), u(t-2), \dots, u(t-d_u)) \quad (3.10)$$

NARX Ağları paralel ya da seri-paralel mimaride tasarlanabilmektedirler. Paralel NARX mimarisinde çıkış değişken değeri tekrar ağın girişine gönderilmekte ve Şekil 3.2’de görüldüğü üzere, doğrusal olmayan dinamik sistemin tahmini tekrar giriş katmanına verilerek döngü mekanizması kurulmuş olmaktadır. Burada tahmin edilmesi istenilen hedef değişkenin önceki zaman serisi gerçek değerleri ele alınmamaktadır.

Seri-Paralel NARX Mimarisinde ise, Şekil 3.3’te görüldüğü üzere hedef değişkenin önceki zaman serisi gerçek değerleri ağın eğitilmesi sırasında girişe gönderilmektedir. Bu nedenle Seri-Paralel NARX Mimarisinde ağa giriş olarak verilen değerler birbirleriyle daha tutarlıdır (Chatterjee ve diğ, 2012).

Gizli katmandaki j . nöronun çıkış değeri H_j denklem 3.11'deki gibi ifade edilir. Gizli katman'da tanjant hiperbolik aktivasyon fonksiyonu kullanılmıştır. w_{ji} gizli katmanda bulunan j . nöron ve i . giriş nöronu arasındaki ağırlığı göstermektedir.

$$H_j(t) = \tanh(\sum_{i=0}^{d_y} w_{ji}y(t-i) + \sum_{i=0}^{d_u} w_{ji}u(t-i) + b_j) \quad (3.11)$$

C gizli katmandaki nöron sayısı, w_{oj} gizli katmandaki j . nöron ve çıkış katman nöronu o arasındaki ağırlığı göstermek üzere çıkış değeri denklem 3.12'deki gibi ifade edilir.

$$\hat{y}(t+1) = \sum_{j=1}^C w_{oj}H_j(t) + b_o \quad (3.12)$$

Her iki mimaride de sinir ağı performansını ölçmek için Ortalama Karesel Hata (MSE - Mean Square Error) kullanılmakta ve MSE değeri 0'a yakın modeller daha başarılı performans göstermektedir. Gizli katmandaki nöron sayısı deneme yanılma yöntemiyle, en düşük MSE değerini verecek şekilde optimize edilmeye çalışılmaktadır (Ardalani ve Zolfaghari, 2010).

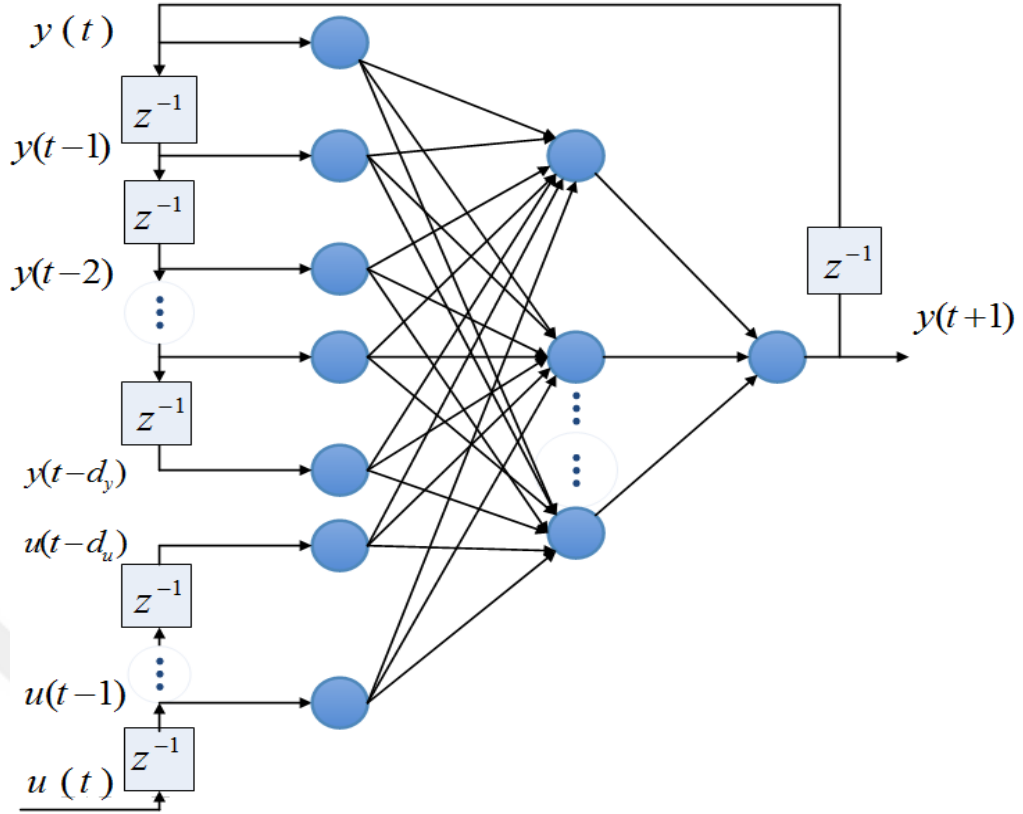
3.4 Kullanılan Araç ve Yazılımlar

3.4.1 UJMP – Universal Java Matrix Package

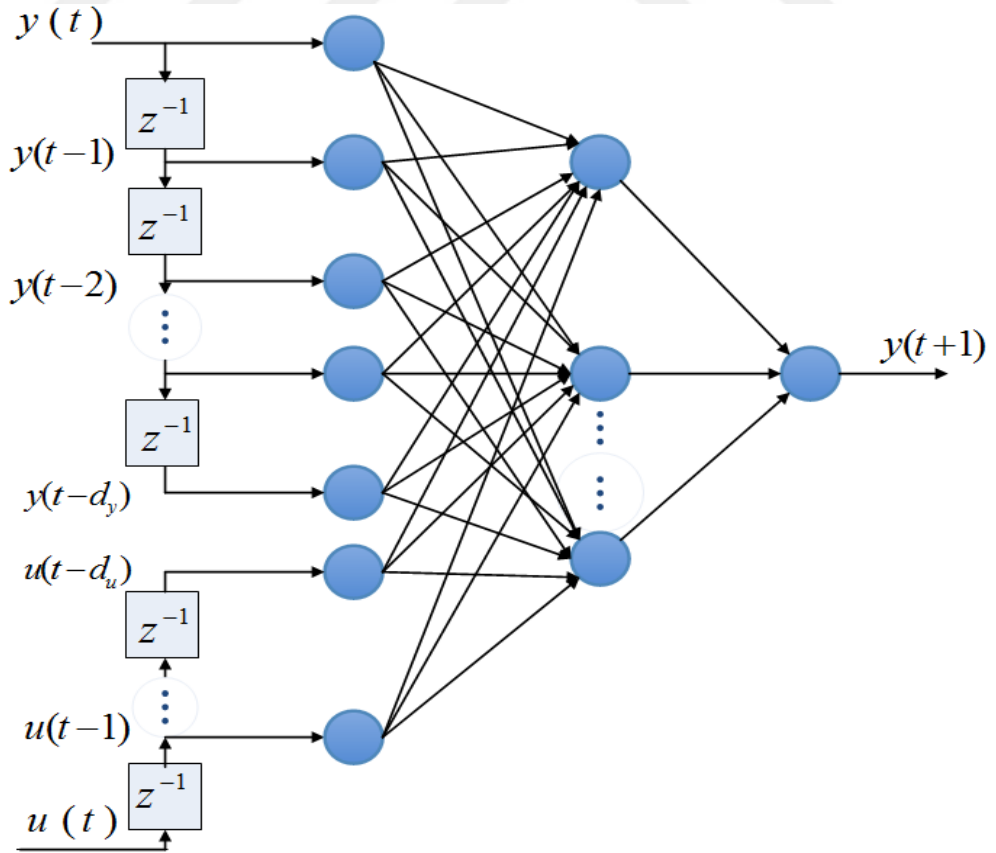
Tez çalışması kapsamında seyrek komşuluk matrislerinin temsili için UJMP (Url-1)'den yararlanılmıştır. UJMP Java kütüphanesi birçok matris işlemlerini gerçekleştirmekte ve seyrek ve yoğun matrisler için lineer cebir hesaplamalarına destek vermektedir. UJMP farklı veri tipleri ve büyük boyutlu matris işlemleri için stabil yapı sunmaktadır.

3.4.2 Stax – Streaming XML for JAVA

StAX (Streaming API for XML) (Url-2), XML formatındaki dokümanlara yazma ve okuma işlemleri için kullanılan bir Java kütüphanesidir. XML API'leri DOM tabanlı yöntemler ve olay (event) tabanlı yöntemler olmak üzere ikiye ayrılmaktadır.



Şekil 3.2 : Paralel NARX mimarisi (Chatterjee ve diğ, 2012).



Şekil 3.3 : Seri-Paralel NARX mimarisi (Chatterjee ve diğ, 2012).

StAX ayrıştırıcı xml dosyaları üzerinde işlem yaparken belleğe tüm dosya yüklenmediği için daha verimli çalışır. StAX API'si java 1.6 ile desteklenen hem xml okuyan hemde xml yazabilen kapsamlı bir programlama arayüzüdür. StAX yöntemini kullanan uygulamalar, çalışma sırasında karşılaştıkları her etiket tanımı için bir olay (event) oluştururlar. Uygulamada tanımlanan olay yöneticileri (event-handler) bu olaylar için gereken işlemleri yaparlar. Sorumluluğun uygulamada olması nedeniyle StAX ayrıştırıcısının xml dosyasının tamamını belleğe yüklemesi gerekmez.

Tez çalışmasında, yüksek boyutlu (örneğin 3 GB) xml dosyalarının bağlantılı liste veri yapılarına okunabilmesi için StAX kütüphanesi kullanılmıştır. StAX ile, xml dosyası içindeki bilgilerin parça parça belleğe alınarak işlenmesi ve anlamlı hale getirilmesi gerçekleştirilmiştir. Xml dosyasındaki tüm etiketler sırayla bellekteki “düğüm” ve “bağlantı” türündeki nesnelere StAX kullanılarak okunmuştur. Düğüm ve bağlantı türündeki nesneler oluşturulduktan sonra, düğüm ve bağlantı nesneleri arasında gerekli ilişkiler kurularak çizge yapısı bağlantılı liste olarak bellekte temsil edilmiştir. Çok büyük verilerin bellekteki temsili ve işlenmesi zor olacağı için matris yapısı tercih edilmemiştir. Çalışmada ağdaki komşuluk ilişkileri, hesaplanması istenen benzerlik ölçütüne göre, “ağırlıklı” veya “ağırlıksız” şekilde belirlenen bağlantılı listelere okunmuştur.

3.4.3 R – İstatistiksel Yazılım Geliştirme Ortamı

R istatistiksel yazılım geliştirme platformu (Url-3) istatistiksel hesaplamalar ve tahmin yapılması için ortam sunmaktadır. R dilinin sentaksı C diline benzerlik göstermekte ve efektif veri işleme ve saklama özelliği sunmaktadır. Dizi ve özellikle matris hesaplamalarında kullanılabilecek özel operatörler mevcuttur. Tek değişkenli ve çok değişkenli zaman serisi analizi için kullanılabilecek uyumlu ve bir arada kullanılabilen araçlar içermektedir (Özdemir ve diğ, 2010). R, belleğe direkt erişim yerine özel veri yapılarını kullanmaktadır. R'deki temel nesne türleri şu şekilde ifade edilmektedir.

- *Vektörler:* R' de altı farklı temel vektör tipi bulunmaktadır: logical, integer, real, complex, string ve raw.
- *Listeler:* Listeler de vektördür ancak listedeki elemanlar farklı tiplerde olabilirler.

- *İfade*: Bir veya daha fazla deyimden oluşan nesnelerdir.
- *Ortamlar*: Sembol-değer çiftlerini içeren bir çerçeveden meydana gelir.

R fonksiyonları ayrı paketler halinde düzenlenmişlerdir. Böylece gerekli paketlerle çalışarak daha az bellek kullanımı ve hızlı işlem gücü sağlanır. Bu paketlerin bir başka avantajı da yazılan fonksiyonlardan oluşan paketlerin R web sitesinden temin edilerek yüklenebilmesidir.





4. ÖNERİLEN YÖNTEMLER

Tez çalışmasının temel amacını oluşturan homojen ve heterojen sosyal ağlar üzerinde bağlantı tahmini problemine çözüm getirebilmek için yeni yöntemler önerilmiştir. Önerilen yöntemler, zaman bilgisinin etkin şekilde kullanılması ve zaman içinde değişim bilgisinin uygun şekilde temsil edilerek mevcut yöntemlerin başarımının artırılmasını amaçlamaktadır. Öncelikle bağlantı tahmini için çok değişkenli zaman serileri tabanlı bir yöntem önerilmiştir. Ardından zaman serisi modelleri gibi serinin durağan olmasını ve serinin doğrusal bir süreçten oluştuğunu önşart olarak kabul etmeyen NARX sinir ağı modeli tabanlı bir yöntem önerilmiştir. İlgili yöntemler; Bölüm 4.2 ve Bölüm 4.3'te detaylı olarak açıklanmıştır.

4.1 Kullanılan Topolojik Tabanlı Benzerlik Ölçütleri

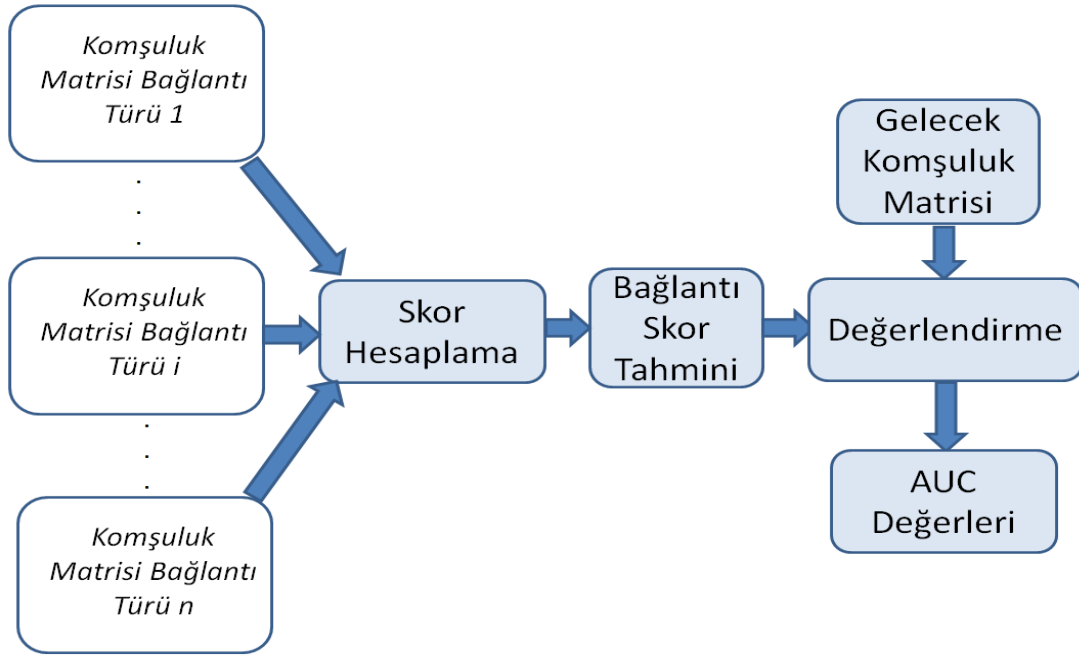
Bağlantı tahmini için önerilen topolojik bilgiyi kullanan benzerlik ölçütlerindeki temel düşünce, $\Gamma(u)$ ve $\Gamma(v)$ kesişim kümesi ne kadar büyük olursa u ve v düğüm çiftleri o kadar yüksek olasılıkla ilerideki zaman diliminde bağlanacaktır. Benzerlik skorları “ağırlıklı” ve “ağırlıksız” olarak iki farklı şekilde hesaplanmıştır. Ağırlıklı benzerlik ölçüt skorlarının hesaplanmasında düğümler arası gerçekleşen bağlantı sayısı göz önünde bulundurulurken, ağırlıksız skorlarda ise düğümler arası bağlantı olup olmadığı dikkate alınmıştır. Lokal (komşuluk tabanlı) ölçütler ve global (tüm yolların birleşimini temel alan) ölçütlerin ilgili skor formülleri Bölüm 2.1.1'de verilmiştir.

4.2 Çok Değişkenli Zaman Serileri ile Bağlantı Tahmini

Tez çalışması kapsamında önerilen ilk yöntem olan Çok Değişkenli Zaman Serileri ile Bağlantı Tahmini'nin anlam ve önemini, zamanla değişen homojen/heterojen sosyal ağlar üzerinde bağlantı tahmini sorununa çözüm getirebilmek için zaman bilgisinin etkin şekilde kullanılması ve zaman içinde değişim bilgisinin uygun şekilde temsil edilerek mevcut yöntemlerin başarımının artırılması oluşturmaktadır.

Önerilen yöntem ile zaman bilgisinin etkin şekilde kullanılması, tahmin edilecek bağlantının gözlenmiş olması kabulünün ortadan kaldırılması, düğüm çiftleri arasındaki bağlantıların ve benzerlik ölçütlerinin korelasyonunun kullanılması gerçekleşmiş ve mevcut yöntemlerin eksikliklerinin giderilmesi sağlanmıştır. Bağlantı tahmini için VAR modeline girdi olarak benzerlik skorları ve bağlantı sıklık bilgisi verilerek eş zamanlı olarak ilişkiler ele alınmıştır.

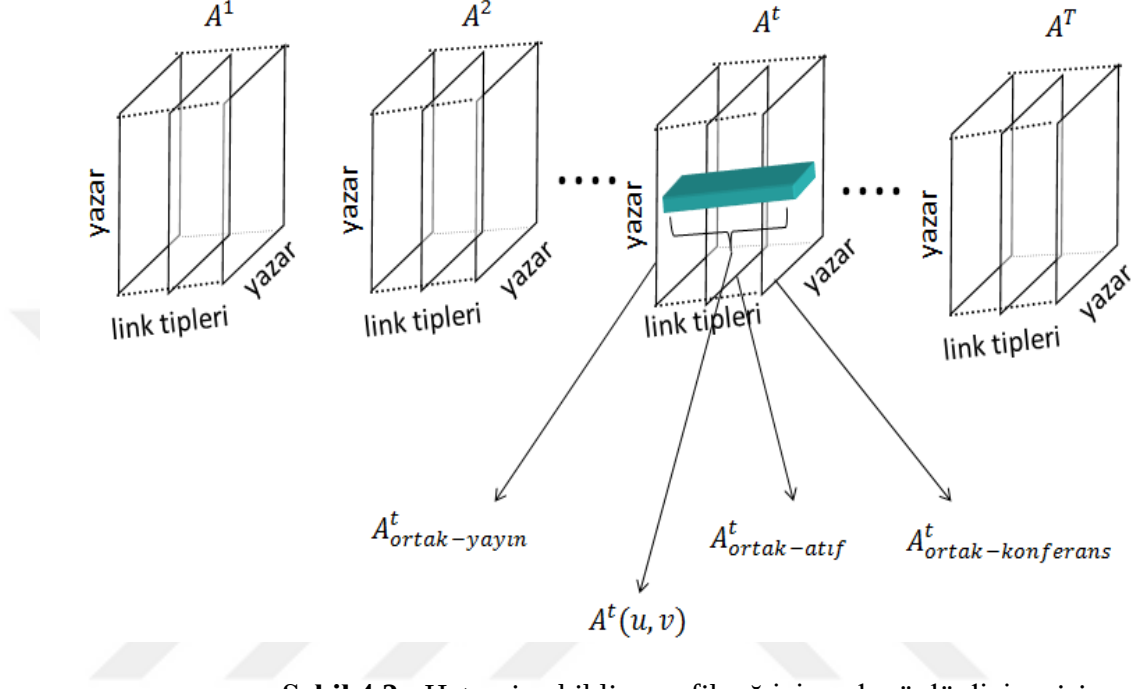
Evrimsel ağlarda bağlantı tahmini için, geçmişte meydana gelen bağlantı sıklık bilgisinin ve benzerlik ölçütlerinin (PropFlow, CN, AA, PA v.s) zaman içinde değişimleri eşanlı bir şekilde T zaman dilimi boyunca analiz edilerek, $T+1$ zaman dilimindeki değerlerinin (çok değişkenli zaman serileri ile) tahmin edilmesi gerçekleşmiştir. Tahmin edilen benzerlik ölçütlerinin skor değerleri, düğüm çiftleri arasındaki bağlantıların tahmini için kullanılmaktadır. Yöntemin adımları Şekil 4.1’de verilmiştir.



Şekil 4.1 : Önerilen yöntemin aşamaları.

Önerilen Çok Değişkenli Zaman Serileri ile Bağlantı Tahmini yöntemi öncelikle homojen ağlara uyarlanmış ardından aynı tür düğümler ve farklı tür bağlantılardan oluşan heterojen ağlara uyarlanmıştır.

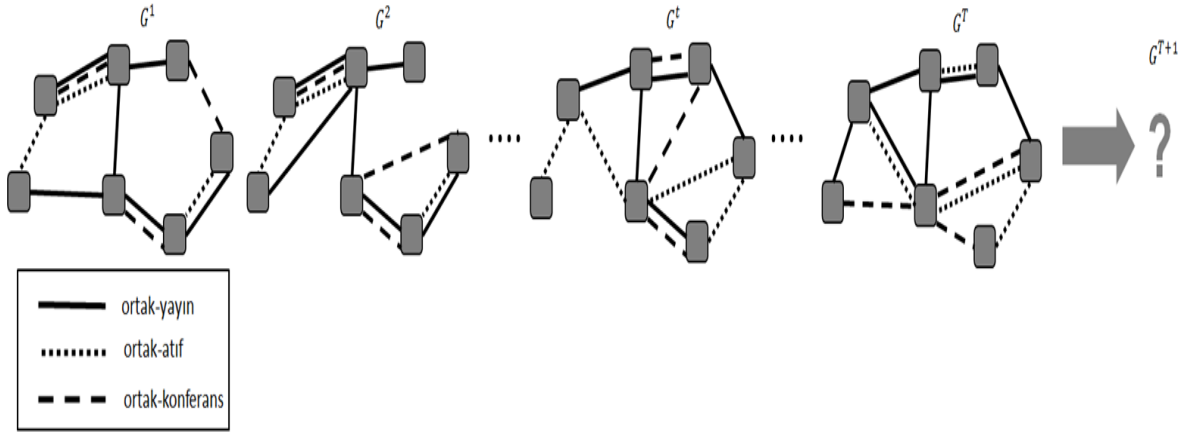
Oluşturulan evrimsel heterojen ağ, zamana göre değişen ağ serisi $(G^1, \dots, G^t, \dots, G^T)$ şeklinde temsil edilmiştir. Ağın her bir görüntüsü $ixjk$ boyutlu, çok yönlü dizi olarak temsil edilmiştir. Burada i ve j düğümleri temsil ederken k ise bağlantı türünü temsil eder. k düğümler arası farklı bağlantı türleri olmak üzere evrimsel heterojen ağ $(A^1, \dots, A^t, \dots, A^T)$ şeklinde çok yönlü diziler serisi ile temsil edilmiştir (Şekil 4.2).



Şekil 4.2 : Heterojen bibliyografik ağ için çok yönlü dizi serisi.

Örnek olarak 3 farklı bağlantı türünden oluşan evrimsel heterojen akademik yayın işbirliği ağı Şekil 4.3'te gösterilmiştir. Buradaki bağlantı türleri ortak yayın yapma ilişkisi (ortak-yayın), ortak yayınlara atıfta bulunma ilişkisi (ortak-atıf) ve ortak konferanslarda yada makalelerde yayın yapma ilişkisi (ortak-konferans) olarak belirlenmiştir.

Her bağlantı (u,v) için (A_k^t) değeri t . zaman dilimindeki k . bağlantı türü için yönsüz bağlantı (u,v) 'nin meydana gelme sıklığını göstermektedir. $(A_k^1(u, v), \dots, A_k^t(u, v), \dots, A_k^T(u, v))$ k . bağlantı türü için bağlantıların meydana gelme sıklığının tek değişkenli zaman serisini vermektedir.



Şekil 4.3 : Evrimsel heterojen bibliyografik ağ örneği.

Her zaman dilimi için tanımlanan çok yönlü diziler kullanılarak, her bağlantı türü için bağlantı tahmini benzerlik ölçüt skorları hesaplanmıştır. Oluşturulan VAR modelinde, değişkenler durağan, hata terimleri ise beyaz gürültü sürecine sahiptir. Modelde her bir değişken; sistemdeki kendi gecikmeli değerleri ve sistemdeki diğer tüm değişkenlerin gecikmeli değerlerinin doğrusal bir fonksiyonudur.

Ayrıca değişkenler arasında geri besleme ilişkisi vardır. Önerilen yöntemin önemli bir avantajı, bağlantının doğrudan meydana gelme sıklığının benzerlik ölçüt skorları ile birlikte kullanılması ve bağlantı türleri arasındaki zaman içindeki korelasyonun göz önünde bulundurulmasıdır. ARIMA/ARMA gibi tek değişkenli zaman serisi modellerinin eksikliğini gidermek için VAR çok değişkenli zaman serisi modeli kullanılmıştır.

Teorik olarak ideal modelde eş zamanlı olarak, tüm bağlantılar arasındaki korelasyonlar düşünülmelidir. Fakat tüm bağlantılar arasındaki korelasyonların hesaplama maliyetinin çok fazla olacağı nedeniyle u , v düğüm çifti arasındaki bağlantı türleri arasındaki korelasyonun eş zamanlı düşünüldüğü, u , v düğüm çifti arasındaki bireysel bağlantıların zaman içinde değişen bağımlılıkların dikkate alındığı ve benzerlik ölçütleri ile birleştiren çok değişkenli zaman serisi tabanlı yöntem önerilmiştir.

Önerilen modelde ağdaki değişim, yeni düğüm eklenmesi ve çıkarılması şeklinde değil, bağlantı ekleme veya çıkarma şeklinde ele alınmıştır. Düğüm kümesinin sabit olduğu farzedilmiştir.

Gerçeklenen yöntemler şunlardır:

➤ Çok değişkenli zaman serileri

- *VAR-Sim Modeli (Vector Autoregressive Model using similarity measures and link occurrences)*: VAR modeli değişkenler arasında dinamik ilişkileri verebilen ve böylelikle zaman serileri alanında değişkenler arasındaki ilişkileri ve geleceğe yönelik tahminleri açıklayabilen bir yöntemdir. Benzerlik ölçüt skorları ve düğüm çiftleri arasındaki bağlantı bilgisi zaman serisi modeline girdi olarak kullanılmıştır.

a. Ağırlıksız : Zaman serilerine girdi olarak bütün bağlantı türleri için ağırlıksız düğüm benzerlik skorları ve düğüm çiftleri arasındaki bağlantı bilgisi kullanılır. Herhangi iki düğüm çifti bir zaman dilimi içinde en az bir defa bağlantı gerçekleştirmiş ise “bağlı”, bağlantı gerçekleştirmemiş ise “bağlı değil” olarak kabul edilmekte ve bu duruma bağlı olarak benzerlik skorları hesaplanmaktadır.

- ✓ AA çok değişkenli zaman serileri (ağırlıksız)
- ✓ CN çok değişkenli zaman serileri (ağırlıksız)
- ✓ JA çok değişkenli zaman serileri (ağırlıksız)
- ✓ PA çok değişkenli zaman serileri (ağırlıksız)
- ✓ Katz çok değişkenli zaman serileri (ağırlıksız)
- ✓ PropFlow çok değişkenli zaman serileri (ağırlıksız)

b. Ağırlıklı : Düğüm benzerlik skorları olarak ağırlıklı skorlar kullanılır. İki düğümün sadece bağlı olmasından ziyade, bağlantıların kaç defa gerçekleştiği bilgisi de göz önünde bulundurulur.

- ✓ AA çok değişkenli zaman serileri (ağırlıklı)
- ✓ CN çok değişkenli zaman serileri (ağırlıklı)
- ✓ JA çok değişkenli zaman serileri (ağırlıklı)
- ✓ PA çok değişkenli zaman serileri (ağırlıklı)
- ✓ Katz çok değişkenli zaman serileri (ağırlıklı)
- ✓ PropFlow çok değişkenli zaman serileri (ağırlıklı)

- *VAR-AllSim Modeli (Vector Autoregressive Model using all similarity measures and link occurrences)*: Başarılı sonuçlar veren ölçütler ile düğüm çiftleri arasındaki bağlantı bilgisi zaman serisi modeline girdi olarak kullanılmıştır.

➤ Tek değişkenli zaman serileri

- *ARIMA-Sim Modeli*: Düğüm benzerlik skorları girdi olarak kullanılmıştır.
- *ARIMA-Occ Modeli*: Bağlantı sıklık bilgileri girdi olarak kullanılmıştır.

4.3 NARX ile Bağlantı Tahmini

Zaman serisi modellerinin önşart olarak ele aldığı varsayımlarından en önemlileri serinin durağan olması ve doğrusal olmasıdır. Bir zaman serisinin durağan olması, zaman içinde belirli bir değere doğru yaklaşması veya beklenen değeri etrafında dalgalandığını ifade etmektedir. Zaman serisi modellerindeki bu eksikliği gidermek için seri ile ilgili herhangi bir varsayımda bulunmayan NARX Sinir Ağı modeli kullanılmıştır.

Bu kapsamda gerçekleştirilen yöntemler şunlardır:

➤ NARX Modelleri

- *NARX-Sim Modeli (NARX model using similarity measures)*: NARX-Sim modeli değişkenler arasında dinamik ilişkileri verebilen ve böylelikle zaman serileri alanında değişkenler arasındaki ilişkileri ve geleceğe yönelik tahminleri açıklayabilen bir yöntemdir. İlgili bağlantı türü için benzerlik skorları girdi olarak kullanılmıştır dışsal değişken kullanılmamıştır.
- *NARX-Sim-TE Modeli (NARX model using similarity measures and temporal evolution information)*: İlgili bağlantı türü için benzerlik skorları ve bağlantı sıklık bilgisindeki değişimler girdi olarak kullanılmıştır. Bağlantı sıklık bilgisindeki değişimler dışsal değişken olarak kullanılmıştır.

- *NARX-Sim-TE-Het Modeli (NARX model using similarity measures and temporal evolution information for heterogeneous network)*: Bütün bağlantı türleri için ilgili benzerlik skorları ve bağlantı sıklık bilgisindeki değişimler girdi olarak kullanılmıştır. Tahmin edilecek bağlantı türü dışındaki diğer bağlantı türlerine ait sıklık bilgisi değişimleri ve benzerlik skorları dışsal değişken olarak kullanılmıştır.

Benzerlik ölçülerine ek olarak bağlantı tahmininde giriş değeri olarak denklem 4.1, 4.2 ve 4.3'te tanımları verilen bağlantı sıklık bilgisindeki değişim bilgileri kullanılmıştır.

$$C(u, v)_j^t = \begin{cases} w(u, v)_j^{t+1} - w(u, v)_j^t, & \text{eğer } (u, v) \in E_j^{t+1}, w(u, v)_j^{t+1} > w(u, v)_j^t \\ 0, & \text{diğer} \end{cases} \quad (4.1)$$

$$P(u, v)_j^t = \begin{cases} \min(w(u, v)_j^t, w(u, v)_j^{t+1}), & \text{eğer } (u, v) \in E_j^t \cap E_j^{t+1} \\ 0, & \text{diğer} \end{cases} \quad (4.2)$$

$$R(u, v)_j^t = \begin{cases} w(u, v)_j^t - w(u, v)_j^{t+1}, & \text{eğer } (u, v) \in E_j^t, w(u, v)_j^t > w(u, v)_j^{t+1} \\ 0, & \text{diğer} \end{cases} \quad (4.3)$$

E_j^t ve E_j^{t+1} kümeleri u ve v düğüm çifti arasında bulunan ardışık iki zaman aralığındaki j ilişki türüne ait bağlantı kümelerini temsil etmek üzere, zamansal bağlantı sıklık bilgisi değişimleri şu şekilde hesaplanmaktadır. u ve v bağlı düğüm çifti arasındaki bağlantının ağırlığının azalması durumunda “kopan bağlantı sayısı” $R(u, v)_j^t$, bağlantının ağırlığının artması durumunda “oluşan bağlantı sayısı” $C(u, v)_j^t$, bağlantının ağırlığının korunması durumunda “korunan bağlantı sayısı” $P(u, v)_j^t$, bağlantı türü j için hesaplanmakta ve bağlantı tahmininde kullanılmaktadır.

➤ Tek değişkenli zaman serileri

- *ARIMA-Sim Modeli*: Düğüm benzerlik skorları girdi olarak kullanılmıştır.

- *ARIMA-Occ Modeli*: Bağlantı sıklık bilgileri girdi olarak kullanılmıştır.

Örneğin, düğümlerinin yazarları temsil ettiği ve “ortak yayın yapma”, “ortak atıfta bulunma”, “ortak konferansta yayın yapma” bağlantı türlerinin temsil ettiği heterojen bir ağda herhangi bir u, v düğüm çifti için $CN_{ortak-yayın}^t$, $R(u, v)_{ortak-yayın}^t$, $C(u, v)_{ortak-yayın}^t$, $P(u, v)_{ortak-yayın}^t$ sırasıyla; t zamanındaki ortak-yayın bağlantı türüne ait benzerlik ölçüt skorunu, kopan bağlantı sayısını, oluşan bağlantı sayısını ve korunan bağlantı sayısını göstermek üzere u, v arasındaki ortak yayın bağlantı sayısı hem d_u dönem önceki dönemlerin ortak yayın türündeki bağlantı sıklık değişim bilgilerine hem de d_y dönem önceki dönemlerin ortak yayın benzerlik ölçüt skorlarına ayrıca d_u dönem önceki dönemlerin ortak atıf ve ortak konferans bağlantı sıklık değişimi ve benzerlik ölçüt skorlarına bağlıdır. Yöntemin önemli bir avantajı, bağlantı değişim bilgisinin benzerlik ölçütleri ile birlikte kullanılması ve bağlantı türleri arasındaki zaman içindeki korelasyonun göz önünde bulundurulmasıdır. ARIMA/ARMA gibi tek değişkenli zaman serisi modellerinin eksikliğini gidermek için NARX çok değişkenli sinir ağı modeli kullanılmıştır.

5. DENEYSEL SONUÇLAR

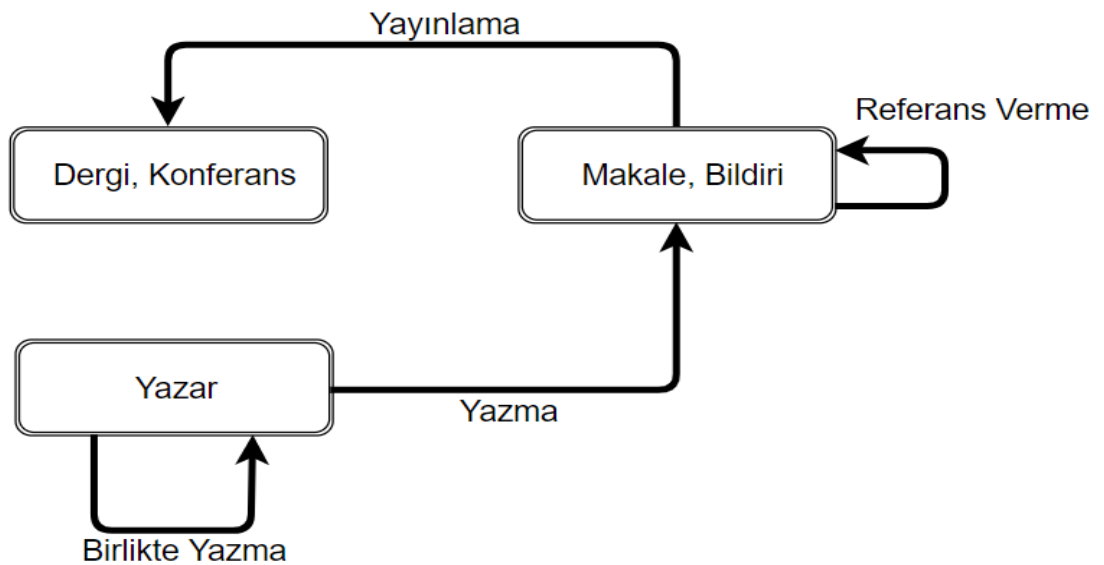
Bu bölümde öncelikle kullanılan veri kümeleri hakkında bilgi verilmiş ardından test metodolojisi açıklanmıştır. Geliştirilen yöntemlerin başarımı, düğümlerin farklı zamanlardaki bağıllık koşulları altında ölçülmüş ve literatürde önerilen diğer yöntemlerle karşılaştırılmıştır. Son olarak detaylı test sonuçları açıklanmıştır.

5.1 Kullanılan Veri Kümeleri

Zamana göre değişen ve farklı bağlantı türlerinden oluşan iki farklı heterojen veri kümesi kullanılmıştır. Veri kümeleri, bilgisayar bilimleri alanında bibliyografik veri kümesi ve kullanıcıların sık kullandığı internet sitelerini sakladığı sosyal sık kullanılan bilgilerini içeren veri kümesinden oluşmaktadır.

5.1.1 DBLP bilgisayar bilimleri bibliyografik veri kümesi

Çalışma için kullanılan veri kümelerinden ilki DBLP Bibliyografik ağ sitesinden (Url-4) toplanılan DBLP Bilgisayar Bilimleri Bibliyografik veri kümesidir. Veri kümesi 1989 ile 2013 yılları arasında bilgisayar bilimleri alanında yayınlanan makaleler, yazarlar ve konferanslara ilişkin bilgileri içermektedir. Şekil 5.1’de veri kümesindeki düğümler ve ilişkiler görülmektedir.

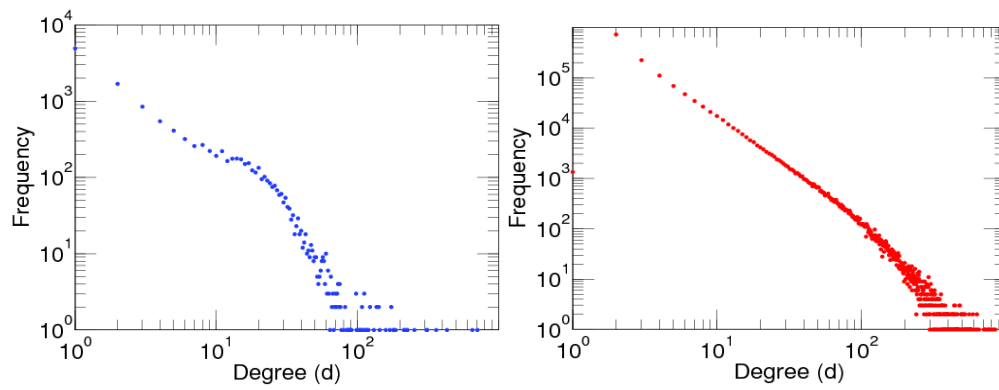


Şekil 5.1 : DBLP veri kümesi yapısı.

Veri kümesi farklı türden bibliyografik ilişkilerle temsil edilmiştir. Veri kümesinde bütün bağlantı türleri için derece dağılımı power-law derece dağılımı özelliğini sergilemektedir. Düğüm ve bağlantılara ait bilgiler Çizelge 5.1’de, atıf verme ve yazma bağlantı türlerine ait power-law derece dağılımı Şekil 5.2’de verilmiştir.

Çizelge 5.1 : DBLP veri kümesi düğüm ve bağlantı türlerine ait ilişki sayıları.

Düğüm/Bağlantı	Düğüm/Bağlantı Türü	Sayı
Düğüm	Yayın(makale, bildiri)	2.092.356
	Yazar	1.712.433
	Dergi, Konferans	22.767
Bağlantı	Referans verme	8.024.869
	Yayın yapma	4.004.883
	Ortak yayın yapma	4.258.615
	Konferans veya dergide yayınlama	1.465.621



Şekil 5.2 : Referans verme ve yazma bağlantı türü derece dağılımı (Url-4-KONECT, 2016).

DBLP veri kümesi üzerinde amaç, makale yazarları arasındaki “ortak yayın yapma”, “ortak atıfta bulunma”, “ortak konferansta yayın yapma” bağlantı türlerinin temsil ettiği heterojen bir ağda bütün ilişki türlerinin tahmin edilmesidir. DBLP veri kümesi üzerinde AUC hesaplanırken, yalnızca çok sayıda bağlantısı olan düğüm çiftleri arasındaki olası bağlantılar dikkate alınmaktadır. Veri ayıklaması aşamasında;

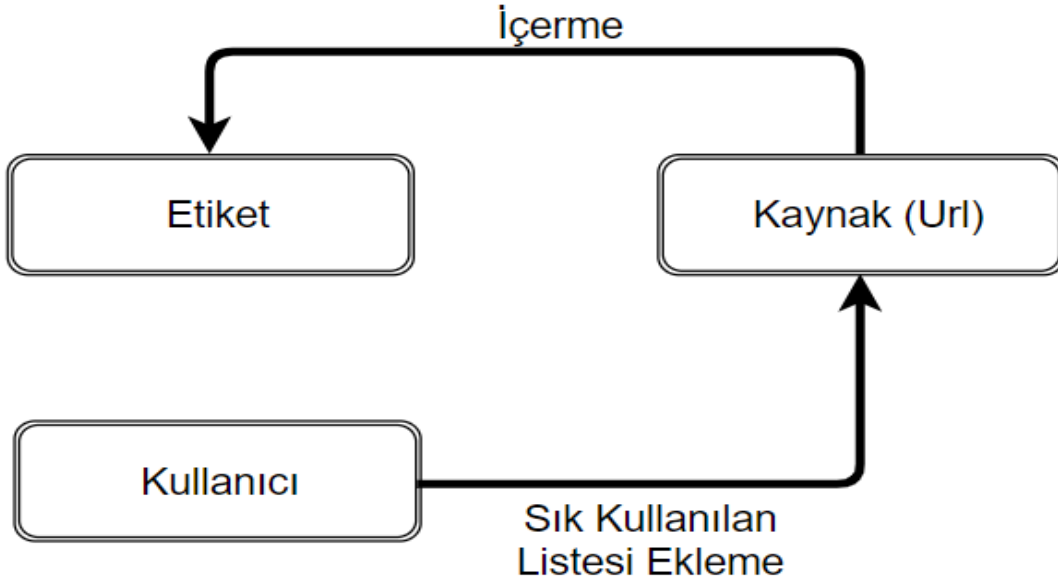
- 2003 yılından önce yayınlanmış bildiriler öncelikle ayıklanmıştır.
- 2003-2012 arasında sık yayın yapan üretken (50’nin üzerinde bağlantısı olan) yazarlar ele alınmıştır. 2003 ile 2012 arasındaki 10 adet 1’er yıllık dönem zaman serilerini oluşturmak için, 2013 yılındaki yayınlar ise test kümesi olarak kullanılmıştır. Bu kısıtlama sonucunda 5931 adet yazar kalmıştır.

5.1.2 Delicious sosyal sık kullanılanlar veri kümesi

Delicious (Url-5), sık kullanılan internet sitesi bağlantılarını depolamak, paylaşmak ve yeni siteleri keşfetmek için kullanılan bir sosyal sık kullanılanlar hizmetidir. Delicious; kullanıcıların beğendiği siteleri kategorize edip takip edebilecekleri ortam sunmaktadır. Delicious herkese açık olup, sık kullanılan linkleri, blogları saklayabilir ve kullanıcılar arasında paylaşımına sunabilmektedir.

Veri kümesi 2005 ile 2011 yılları arasındaki kullanıcılar, kullanıcıların sık kullanılanlara eklemiş olduğu URL’ler (kaynaklar) ve bu kaynaklara vermiş oldukları etiketlere ilişkin bilgileri içermektedir. Şekil 5.3’te veri kümesindeki düğümler ve ilişkiler görülmektedir. Düğüm ve bağlantılara ait bilgiler ise Çizelge 5.2’de verilmiştir. Veri kümesinde bütün bağlantı türleri için derece dağılımı power-law derece dağılımı özelliğini sergilemektedir.

Hesaplama maliyetini uygun hale getirmek için veri kümesi 3’er aylık dönemlere ayrılmış ve 01 Ocak, 2008 ile 30 Eylül 2010 arasındaki 11 adet 3’er aylık dönem zaman serilerini oluşturmak için, 2010 yılının son 3 aylık dönemi ise test kümesi olarak kullanılmıştır. Delicious veri kümesi üzerinde AUC değerleri hesaplanırken, sadece bağlantı sayısı fazla olan düğümler arasındaki olası bağlantılar dikkate alınmaktadır. Hesaplama maliyetlerini uygun duruma getirmek için veri ayıklama aşamasında 20’nin üzerinde kaynağı sık kullanılan listelerine eklemiş olan kullanıcılar seçilmiştir.



Şekil 5.3 : Delicious veri kümesi yapısı.

Çizelge 5.2 : Delicious veri kümesi düğüm ve bağlantı türlerine ait ilişki sayıları.

Düğüm/Bağlantı	Düğüm/Bağlantı Türü	Sayı
Düğüm	Kullanıcı	34.665
	Etiket	16.641
	Kaynak	9.416
Bağlantı	Sık kullanılan ekleme	1.358.522
	Etiketleme	4.546.813

Delicious veri kümesi üzerinde amaç, kullanıcılar arasındaki “sık kullanılanlara ortak kaynakları ekleme” ve “kaynaklar için ortak etiketleri kullanma” bağlantı türlerinin temsil ettiği heterojen bir ağda bütün ilişki türlerinin tahmin edilmesidir.

5.2 Bağlantı Tahmini Test Yöntemleri

Tez kapsamında önerilen yöntemler ile; ağdaki tüm olası bağlantılar için gelecekteki benzerlik skor değerleri tahmin edilerek, bağlantı tahmini yapılmaktadır. Başlangıç olarak, ağdaki olası tüm bağlantılar için hesaplanan skorlar büyükten küçüğe sıralanmakta ve tahmin edilmeye çalışılan gerçek bağlantı bilgileri ile karşılaştırılarak yöntemin başarımı değerlendirilmektedir. Yöntemin başarımı, dengesiz veri kümeleri için de doğru sonuç üreten ve belirli bir eşik değerine ihtiyaç duymadan tüm veri üzerinde hesaplama yapılabilen AUC (Area under ROC curve) cinsinden hesaplanmıştır. AUC hesabı, farklı şekillerde gerçekleştirilmiştir: (Gunes ve diğ, 2015).

- Tekrarlanan ve yeni oluşacak bağlantıları tahmin etmek için tüm olası bağlantılar kullanılarak hesaplanan, (AUC_{all})
- Ağda son 3 zaman diliminde bağlı olmayan düğüm çiftlerinin gelecekteki bağlantı durumlarını tahmin etmek için hesaplanan, ($AUC_{disconn}$). Ağdaki j bağlantı türüne ait t . zaman dilimindeki komşuluk matrisi A_j^t ve tüm düğüm çiftleri (olası bağlantılar) P_j^t için $AUC_{disconn}$ değeri denklem 5.1'deki gibi hesaplanır.

$$AUC_{disconn}(P_j^t) = \{AUC_{all}(p_j^t \subset P_j^t), \forall (u, v) \in p_j^t: A_j^{t-1}(u, v) = 0, A_j^{t-2}(u, v) = 0, A_j^{t-3}(u, v) = 0\} \quad (5.1)$$

- Ağda son 3 zaman diliminde bağlı olan düğüm çiftlerinin gelecekteki bağlantı durumlarını tahmin etmek için hesaplanan, (AUC_{conn}). Ağdaki j bağlantı türüne ait t . zaman dilimindeki komşuluk matrisi A_j^t ve tüm düğüm çiftleri (olası bağlantılar) P_j^t için AUC_{conn} değeri denklem 5.2'deki gibi hesaplanır.

$$AUC_{conn}(P_j^t) = \{AUC_{all}(p_j^t \subset P_j^t), \forall (u, v) \in p_j^t: A_j^{t-1}(u, v) > 0, A_j^{t-2}(u, v) > 0, A_j^{t-3}(u, v) > 0\} \quad (5.2)$$

- Ağda son zaman diliminde bağlı olmayan düğüm çiftleri için hesaplanan, $AUC_{lastDisconn}$.

- Ağda son zaman diliminde bağlı olan düğüm çiftleri için hesaplanan, $AUC_{lastConn}$.
- Ağda son zaman diliminin ardından bağıllık durumunda değişiklik olan (bağlı-bağlı değil, bağlı değil-bağlı) düğüm çiftleri için, AUC_{change} . Denklem 5.3'te ifadesi verilen AUC_{change} 'in tahmin edilmesi diğer ölçümlere göre daha zordur.

$$AUC_{change}(P_j^t) = \{AUC_{all}(p_j^t \subset P_j^t), \forall (u, v) \in p_j^t: (A_j^{t-1}(u, v) > 0, A_j^t(u, v) = 0) \vee (A_j^{t-1}(u, v) = 0, A_j^t(u, v) > 0)\} \quad (5.3)$$

5.3 Test Sonuçları

Bağlantı tahmini testleri kapsamında, evrimsel homojen ve heterojen ağlarda önerilen bağlantı tahmini modelleri ile literatürdeki mevcut zaman bilgisini kullanan modellerin başarımları karşılaştırılmıştır.

Öncelikle veri kümelerinin farklı ardışıl zaman dilimlerindeki görüntüleri oluşturulmuş ve her zaman dilimindeki ağ görüntüsü için düğüm çiftleri arasında çeşitli ağırlıklı ve ağırlıksız benzerlik ölçüt skorları hesaplanmıştır (Murata ve Moriyasu, 2007). Benzerlik ölçütleri olarak Bölüm 4'te tanımları verilen lokal (komşuluk tabanlı) ölçütler, global (tüm yolların birleşimini temel alan) ölçütler, heterojen benzerlik ölçütleri kullanılmıştır. Testlerin ilk aşamasında önerilen çok değişkenli zaman serisi modeli ile evrimsel homojen ağlarda bağlantı tahmini gerçekleştirilmiş ve başarımları tek değişkenli zaman serisi modelleri ile karşılaştırılmıştır.

Test çalışmasının ikinci aşamasında çok değişkenli zaman serisi modeli evrimsel heterojen ağlar için genişletilmiş ve farklı türlere ait bağlantıların tahminlerine ait sonuçlar verilmiştir. Bağlantı tahmini için VAR modeline girdi olarak farklı bağlantı türleri için hesaplanan benzerlik skorları ve bağlantı sıklık bilgisi verilerek zaman içindeki değişimleri eşanlı ele alınmıştır. Testlerin üçüncü aşamasında ise yine evrimsel heterojen ağlar için önerilen NARX sinir ağı modeli tabanlı bağlantı tahmini yönteminin başarımları tek değişkenli zaman serisi modelleri ile karşılaştırılmıştır. NARX modeline girdi olarak farklı bağlantı türleri için hesaplanan

benzerlik skorları ve bağlantı sıklık bilgisindeki değişimler verilerek eş zamanlı olarak ele alınmıştır.

5.3.1 VAR modeli ile homojen ağlarda bağlantı tahmini testleri

Bağlantı tahmini testlerinin temel amacı, önerilen çok değişkenli zaman serileri modeli (VAR-Sim modeli) ile literatürdeki tek değişkenli zaman serileri modellerinin (MA, RW, ARIMA) bağlantı tahmini performanslarını karşılaştırmaktır. VAR-Sim modelinde “Sim” kullanılan benzerlik ölçütünü göstermektedir. VAR-CN ve VAR-WCN sırasıyla ağırlıksız ve ağırlıklı CN benzerlik skorlarının ve komşuluk bilgisinin çok değişkenli zaman serileri modelleridir. VAR-AllSim ve VAR-WAllSim ise sırasıyla ağırlıksız ve ağırlıklı bütün benzerlik skorlarının ve komşuluk bilgisinin çok değişkenli zaman serileri modelleridir. Karşılaştırma amacı ile tek değişkenli zaman serisi modelleri ARIMA-Sim, MA-Sim, RW-Sim kullanılmıştır. Ayrıca literatürdeki bir diğer yöntem olan statik benzerlik skorları yöntemi karşılaştırma amacı ile eklenmiştir. Statik benzerlik skorları sadece son zaman dilimindeki bağlantılar kullanılarak hesaplanmıştır. Benzerlik ölçütleri olarak lokal ölçütler kullanılmış ve tüm ölçütler için hem ağırlıklı hem ağırlıksız versiyonlar test edilmiştir. DBLP veri kümesi üzerinde düğümleri yazarlardan, bağlantıları ortak yayın yapma ilişkisi olan zamana göre değişen homojen ağ meydana getirilmiş ve evrimsel ağın farklı zamanlardaki görüntüleri oluşturulmuştur. Çalışma kapsamında ağ görüntüleri arasında düğümlerin (yazarların) değişmediği sadece bağlantı bilgisinin değiştiği kabul edilmiştir.

Ağda bulunan düğümler arasındaki komşuluk bilgileri, hesaplanacak benzerlik ölçütüne göre ağırlıklı/ağırlıksız şekilde belirlenen bağlantılı listelere okunmuştur. Ağırlıksız benzerlik ölçüt skorları için iki düğümün ilgili zaman diliminde bağlanıp bağlanmadığı tutulurken, ağırlıklı versiyonlar için tekrarlı bağlantıları kullanmak amacıyla ilgili zaman diliminde yapılan bağlantı sayısı saklanmıştır.

Çizelge 5.3 ve Çizelge 5.4’te DBLP veri kümesi üzerinde, önceki zaman dilimlerinde bağlı ve bağlı olmayan bütün düğüm çiftleri gözetilerek hesaplanan (AUC_{all}) bağlantı tahmini sonuçları yer almaktadır. Sonuçlar, kullanılan yöntemlere göre gruplanmıştır. Bu tablolara göre, önerilen VAR modeli (son satır), PA dışındaki tüm benzerlik skorları için, sadece statik benzerlik kullanımından ve sadece benzerlik

skor bilgisinin tek değişkenli zaman serileri ile modellenmesinden daha iyi sonuç vermektedir.

Çizelge 5.3 : DBLP veri kümesi için, AUC_{all} sonuçları (ağırlıksız).

<i>Model</i>	<i>CN</i>	<i>AA</i>	<i>PA</i>	<i>JC</i>	<i>RA</i>	<i>All</i>
Static	0.78	0.79	0.71	0.77	0.79	-
MA	0.80	0.81	0.73	0.78	0.81	-
RW	0.79	0.79	0.72	0.77	0.79	-
ARIMA	0.84	0.85	0.73	0.82	0.85	-
VAR	0.89	0.90	0.72	0.86	0.90	0.92

Çizelge 5.4 : DBLP veri kümesi için, AUC_{all} sonuçları (ağırlıklı).

<i>Model</i>	<i>WCN</i>	<i>WAA</i>	<i>WPA</i>	<i>WJC</i>	<i>WRA</i>	<i>Wall</i>
Static	0.79	0.80	0.72	0.78	0.80	-
MA	0.82	0.82	0.74	0.79	0.83	-
RW	0.80	0.81	0.72	0.78	0.80	-
ARIMA	0.85	0.86	0.75	0.84	0.87	-
VAR	0.90	0.90	0.74	0.88	0.91	0.93

Buna göre, bağlantı tahmini bağlamında, düğüm benzerliklerinin ve bağlantı bilgisinin zaman içinde değişiminin birlikte ele alınması, düğüm benzerliklerinin zaman içinde değişiminden daha belirleyici olduğu iddia edilebilir. Benzerlik ve bağlantı sıklık bilgisinin zaman içinde değişiminin birlikte ele alınması bağlantı tahmini başarımında artış sağlamaktadır. Geliştirilen model ile zaman bilgisini kullanmayan klasik yöntemlerin (AA, JC, PA, CN, RA) en büyük sorunu olan zamanla değişim bilgisini kullanamama ve tekrarlı bağlantıların tahmin edilememe sorunu ile zaman bilgisini kullanan tek değişkenli zaman serisi yöntemlerinin daha önce örneği görülmeyen (meydana gelmeyen) bağlantılar ile ilgili yeterince bilgi edinememe sorunlarına çözüm üretilmiştir. İlgili problemlerin her ikisine birden çözüm üretmek için bağlantı bilgisi ve benzerlik ölçütleri arasında ortaya çıkan zamansal korelasyon bilgisi kullanılmış ve mevcut yöntemlerin başarımı arttırılmıştır. VAR modeline giriş olarak, ağdaki düğümler arası ağırlıklı/ağırlıksız benzerlik skorlarından ve düğümler arası bağlantı değerlerinden oluşan vektörler verilmiştir. VAR-AllSim yöntemine verilen giriş vektörü bütün benzerlik ölçütlerinden (JC, PA, AA, CN, RA) ve bağlantı bilgisinden oluşmakta, VAR-Sim yönteminde ise giriş verilen vektör her bir ilgili benzerlik ölçütü ve bağlantı bilgisinden oluşmaktadır. Önerilen yöntem, en iyi sonucu bütün benzerlik ölçütlerinin birlikte kullanıldığı durumda verirken, en düşük başarımla ise PA skoru ile

elde edilmektedir. Elde edilen sonuçlara göre, benzerlik skorlarının ağırlıklı versiyonlarının ele alınması, tüm modellerin başarımında artış sağlamaktadır. Bu artışın muhtemel nedeni, verilen bir zaman diliminde tekrar eden bağlantıların da dikkate alınmasıdır.

5.3.1.1 Farklı koşullar altında bağlantı tahmini

Bağlantı tahmini yöntemlerinin başarımı, mevcut düğüm çiftlerinin geçmişteki bağlılık durumlarına göre değişiklik göstermektedir (Gunes ve diğ, 2015). Geçmiş zaman dilimlerinde uzun süre bağlı olmayan düğüm çiftleri için tahmin yapmak zorlaşmakta diğer yandan geçmişte uzun süre bağlı olan düğüm çiftlerinin bağlanıp bağlanmayacağını tahmin etmek ise kolaylaşmaktadır.

Farklı koşullar altında bağlantı tahmini başarımını ölçmek için literatürde tanımlanan çeşitli AUC ölçümleri (AUC_{all} , $AUC_{disconn}$, AUC_{conn} , $AUC_{lastDisconn}$, $AUC_{lastConn}$, AUC_{change}) test aşamasında kullanılmıştır. AUC_{all} sonuçlarına göre karşılaştırılan modeller arasında en iyi sonucu veren ARIMA modeli olduğu için farklı koşullar altında bağlantı tahmini başarımını ölçme testlerinde VAR ve ARIMA modelleri kullanılmıştır. Çizelge 5.5'te, tasarlanan ve karşılaştırma amacı ile kullanılan bazı modellerin, değişik AUC ölçümleri türünden sonuçları görülmektedir. Geliştirilen modeller içinde, en iyi sonuçlar AA, CN, RA benzerlik ölçütleri ile elde edildiği için sadece bu benzerlikleri kullanan yöntemlerin sonuçları verilmiştir. Karşılaştırma amacıyla kullanılan modellerin de yüksek başarımları dahil edilmiştir. Çizelge 5.5'e göre, önerilen yöntemler, özellikle de ağırlıklı düğüm benzerliklerinin modellendiği yöntemler (VAR-Wall, VAR-WCN, VAR-WAA, VAR-WRA), bütün AUC tipleri için diğerlerinden daha başarılı sonuç vermektedir. Geliştirilen modellerin, farklı bağlılık durumlarında da kullanılmasının uygun olduğu belirtilebilir.

Geliştirilen VAR modelinin, sadece benzerlik skor bilgisinin tek değişkenli zaman serileri olarak modellendiği ARIMA ile yakın $AUC_{disconn}$ ve $AUC_{lastDisconn}$ sonuçlarının vermesinin nedeni ise son zaman dilimlerinde bağlantıların eksik olması ve VAR modelinin bağlantı bilgisi ile benzerlik skorlarını yakın geçmişte birlikte kullanamamasıdır. Yakın geçmişteki zaman dilimlerinde eksik bağlantı bilgileri içeren düğüm çiftlerinin zaman serisi analizinde VAR modelinin yeterli bilgi

edinmesi mümkün değildir çünkü bu yöntemde bağlantı bilgisi ile benzerlik skorlarının korelasyon bilgisi kullanılmaktadır.

Son 3 zaman diliminde (AUC_{conn}) ve sadece son zaman diliminde ($AUC_{lastConn}$) bağlı olan düğüm çiftleri için testler gerçekleştirilmiştir. Çizelge 5.5'e göre, önerilen yöntem, her iki ölçüme göre diğer yöntemlerden daha başarılı sonuç vermiştir. Buradaki iyileşme önerilen VAR modelinde daha fazladır, bu durumun nedenleri arasında yakın geçmişteki bağlantı bilgilerinin eksik olmaması ve VAR modelinde korelasyon bilgisinin kullanılabilmesi gösterilebilir. Ayrıca, son zaman dilimine göre bağıllık durumunda değişiklik gösteren düğüm çiftleri için de (AUC_{change}) testler yapılmıştır. AUC_{change} ölçümü diğer AUC ölçümlerine göre tahmin edilmesi daha zor bir durum teşkil etmektedir. Önerilen VAR modeli, bu ölçüm için nispeten daha düşük değerler sağlamasına rağmen, ARIMA modeline göre %5-10 arası gelişme sağladığı görülmüştür.

5.3.1.2 Çapraz geçerleme testleri

Önceki bölümde gerçekleştirilen testler, 10-kat çapraz geçerleme ile tekrarlanmıştır. Zaman serilerinin oluşturulduğu (öğrenme kümesi) son zaman dilimleri, 10 eşit parçaya ayrılmış ve her bir testte ilgili parçada bulunan düğüm çiftleri kullanılmıştır. Böylece 10 farklı parçadaki düğüm çiftleri için zaman serileri modelleri oluşturulmuş ve her parça için farklı bağlantı tahmini ve testler yapılmıştır. Bu 10 farklı kümedeki testler sonucunda elde edilen AUC değerlerinin ortalaması ise son test sonucu olarak verilmiştir. DBLP veri kümesi için test sonuçları Çizelge 5.6'da verilmiştir.

Çapraz geçerleme ile elde edilen sonuçlar, daha az öğrenme verisi ile elde edildiği için, Bölüm 5.3.1.1'de elde edilen sonuçlara göre daha düşüktür. Önerilen modelin performans düşüşü, ARIMA modeline göre daha azdır çünkü VAR modeli daha az öğrenme kümesi ile daha başarılı sonuç vermektedir. Sonuçlardaki standart sapma değeri de geliştirilen modelde, diğer yönteme nispeten daha düşüktür. Bunun nedeni de, geliştirilen modelde hem benzerlik skorları hem de bağlantı bilgisinin zaman içinde değişiminin birlikte kullanılması ve veri kümesindeki değişimden daha az etkilenilmesidir.



Çizelge 5.5 : DBLP veri kümesi için sonuçlar.

Önerilen Yöntemler					Karşılaştırılan Yöntemler		
Ölçüm	VAR-WCN	VAR-WAA	VAR-WRA	VAR-WAll	ARIMA-WCN	ARIMA-WAA	ARIMA-WRA
AUC _{all}	0.90	0.90	0.91	0.93	0.85	0.86	0.87
AUC _{disconn}	0.81	0.82	0.82	0.83	0.80	0.81	0.81
AUC _{conn}	0.72	0.74	0.74	0.78	0.66	0.67	0.67
AUC _{lastDisconn}	0.84	0.84	0.85	0.89	0.82	0.83	0.83
AUC _{lastConn}	0.73	0.75	0.75	0.79	0.69	0.70	0.70
AUC _{change}	0.65	0.65	0.66	0.72	0.61	0.62	0.63

Çizelge 5.6 : DBLP veri kümesi için sonuçlar.

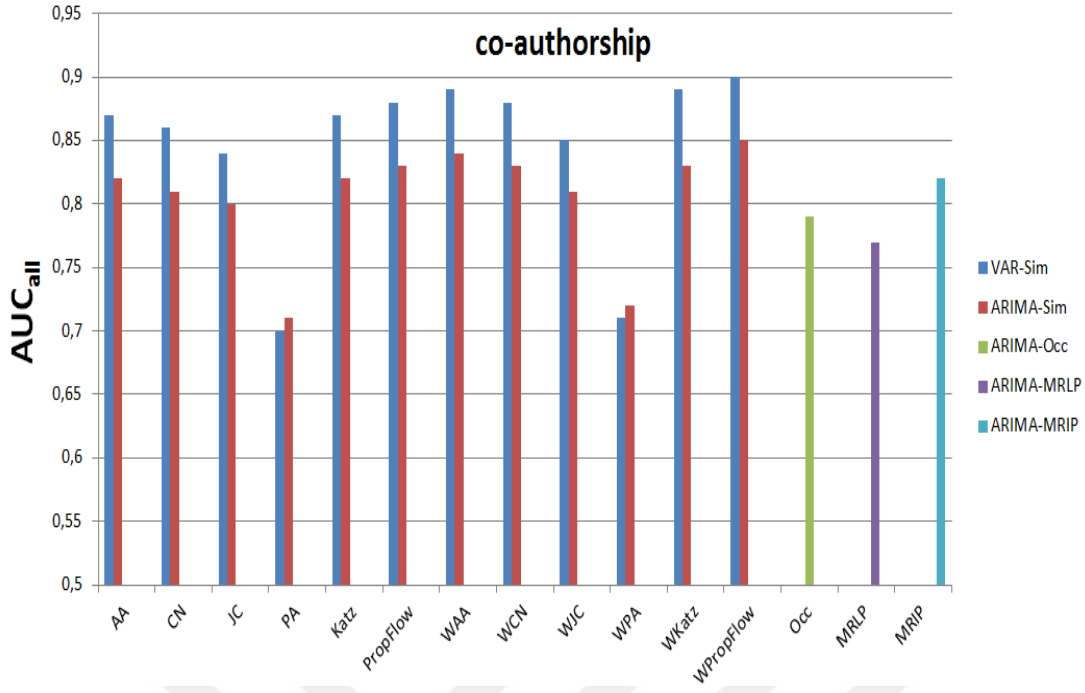
Önerilen Yöntemler					Karşılaştırılan Yöntemler		
Ölçüm	VAR-WCN	VAR-WAA	VAR-WRA	VAR-WAll	ARIMA-WCN	ARIMA-WAA	ARIMA-WRA
AUC _{all}	0.86±0.02	0.86±0.02	0.87±0.01	0.88±0.01	0.80±0.03	0.80±0.03	0.81±0.03
AUC _{disconn}	0.76±0.02	0.76±0.02	0.77±0.01	0.79±0.01	0.75±0.02	0.76±0.03	0.76±0.02
AUC _{conn}	0.69±0.01	0.69±0.01	0.70±0.01	0.74±0.01	0.62±0.02	0.63±0.02	0.63±0.02
AUC _{lastDisconn}	0.79±0.02	0.80±0.02	0.80±0.02	0.84±0.02	0.76±0.02	0.77±0.02	0.78±0.02
AUC _{lastConn}	0.72±0.01	0.72±0.01	0.73±0.01	0.75±0.01	0.65±0.03	0.66±0.02	0.67±0.02
AUC _{change}	0.63±0.01	0.63±0.01	0.64±0.01	0.68±0.01	0.56±0.04	0.57±0.04	0.57±0.03

Elde edilen sonuçlara göre düğüm benzerlik ölçütlerinin ve bağlantı sıklık bilgilerinin zaman içinde değişimlerinin birlikte eşanlı ele alınması bağlantı tahmini başarımında artış sağlamaktadır. Geliştirilen model en başarılı sonucu bütün benzerlik ölçütlerinin birlikte kullanıldığı durumda (VAR-All) verirken en düşük başarımlı PA metriğinde elde edilmektedir. Test sonuçlarının değerlendirilmesi için modeller ikili karşılaştırılarak “İlişkili ölçümler için eşli t-testi (Paired Samples t-test)” yapılmıştır.

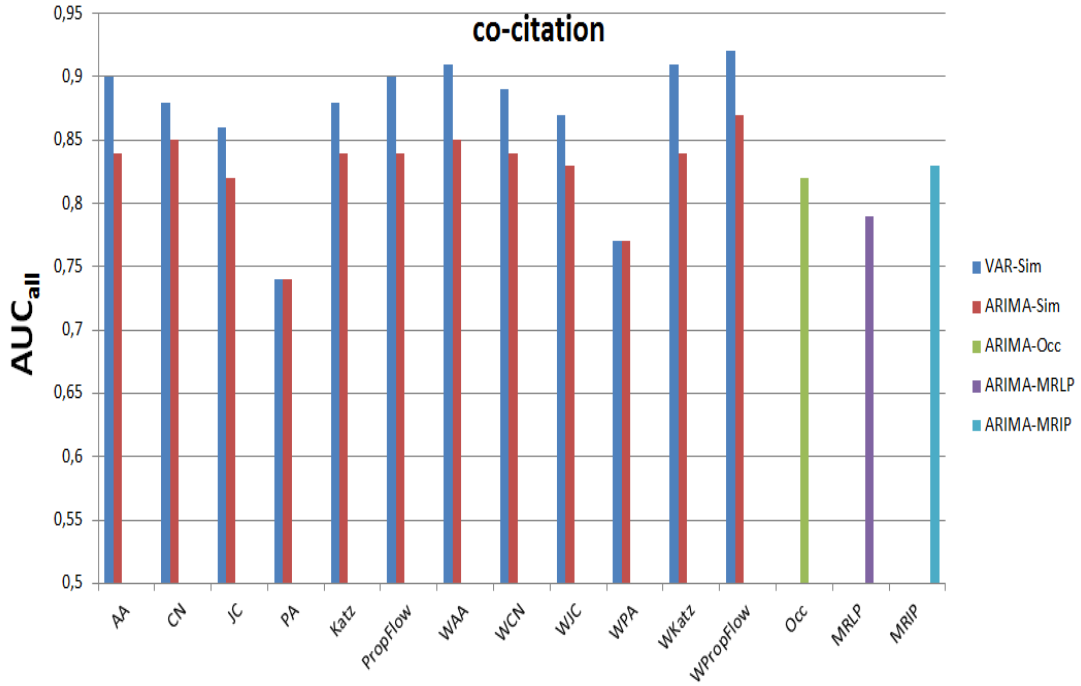
5.3.2 VAR modeli ile heterojen ağlarda bağlantı tahmini testleri

Bağlantı tahmini testlerinin temel amacı, önerilen çok değişkenli zaman serileri modelleri (VAR-Sim ve VAR-AllSim) ile mevcut tek değişkenli zaman serileri modellerinin (ARIMA-Sim ve ARIMA-Occ) bağlantı tahmini performanslarını karşılaştırmaktır. Bu yöntemler arasında, bütün modeller için hem ağırlıklı hem ağırlıksız benzerlik ölçütlerinin kullanıldığı modeller değerlendirilmiştir. VAR-Sim modeline giriş olarak heterojen ağdaki bütün bağlantı türleri için hesaplanmış ilgili benzerlik ölçüt skorları ve düğümler arasında bağlantı sıklık değerlerinden oluşan vektörler verilmiştir. VAR-AllSim modelinde ise başarılı sonuçların elde edildiği AA, PropFlow, Katz ölçüt skorları farklı bağlantı türleri için hesaplanmış ve birden fazla benzerlik ölçüt skorundan oluşan vektörler modele girdi olarak verilmiştir. Çalışma kapsamında ağ görüntüleri arasında düğümlerin değişmediği sadece bağlantı bilgisinin değiştiği kabul edilmiştir.

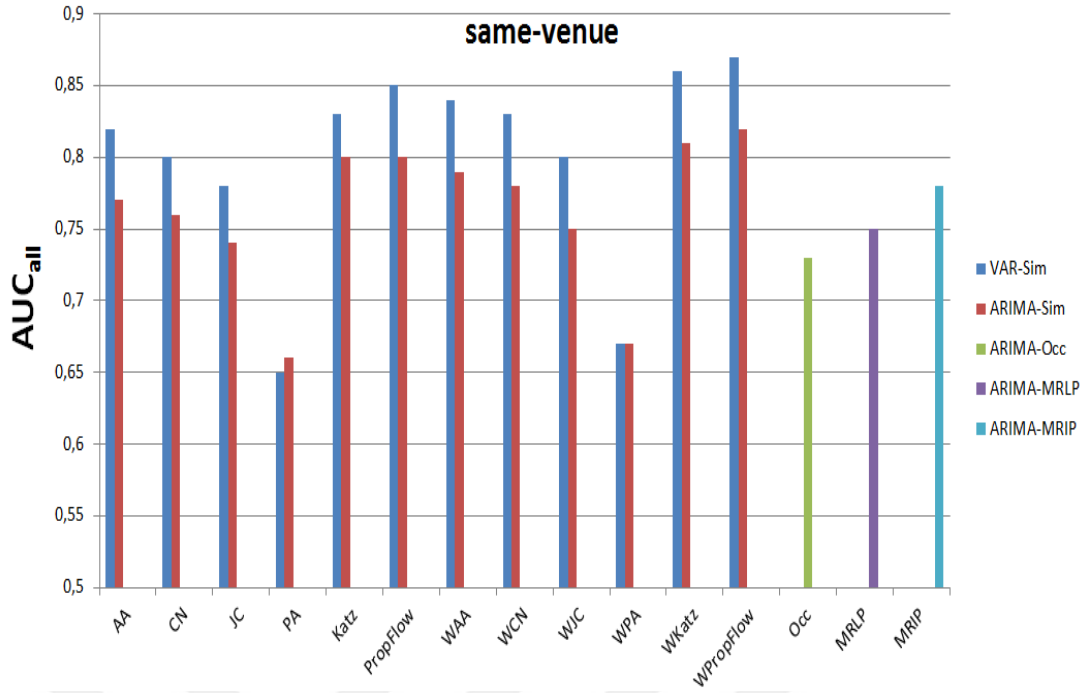
Şekil 5.4, Şekil 5.5 ve Şekil 5.6’da DBLP veri kümesi üzerinde, önceki zaman dilimlerinde bağlı ve bağlı olmayan bütün düğüm çiftleri gözetilerek hesaplanan (AUC_{all}) bağlantı tahmini sonuçları yer almaktadır. Sonuçlar, kullanılan yöntemlere göre gruplanmıştır. Bu sonuçlara göre, önerilen VAR modeli, PA dışındaki tüm benzerlik skorları için, sadece benzerlik skor bilgisinin tek değişkenli zaman serileri ile modellenmesinden ve sadece düğümler arasındaki bağıllık bilgisinin tek değişkenli zaman serileri ile tahmin edilmesinden daha başarılı sonuç vermektedir.



Şekil 5.4 : DBLP veri kümesi “ortak yayın” bağlantı türü için, AUC_{all} sonuçları.



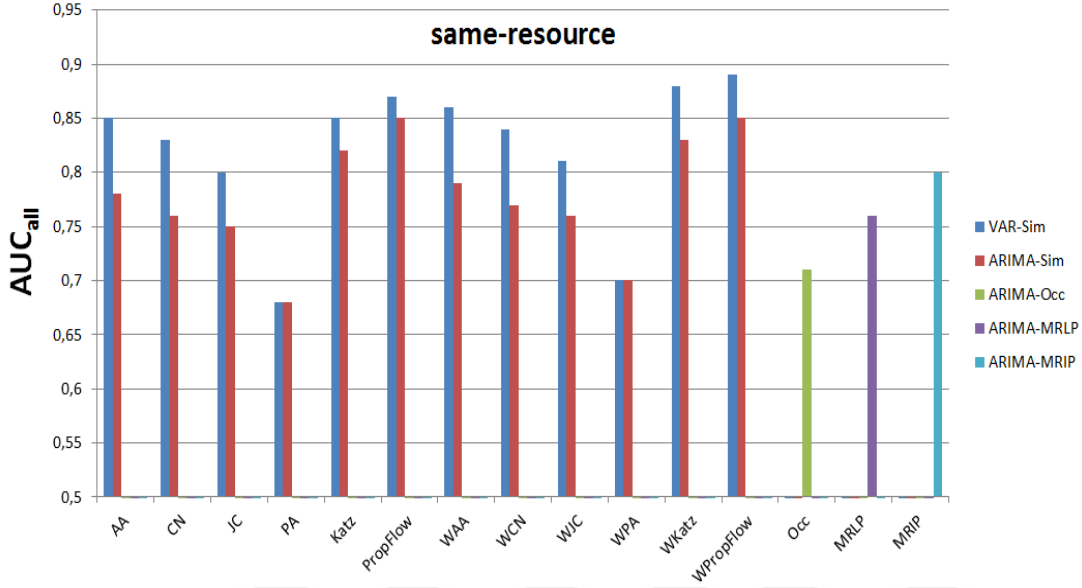
Şekil 5.5 : DBLP veri kümesi “ortak atıf” bağlantı türü için, AUC_{all} sonuçları.



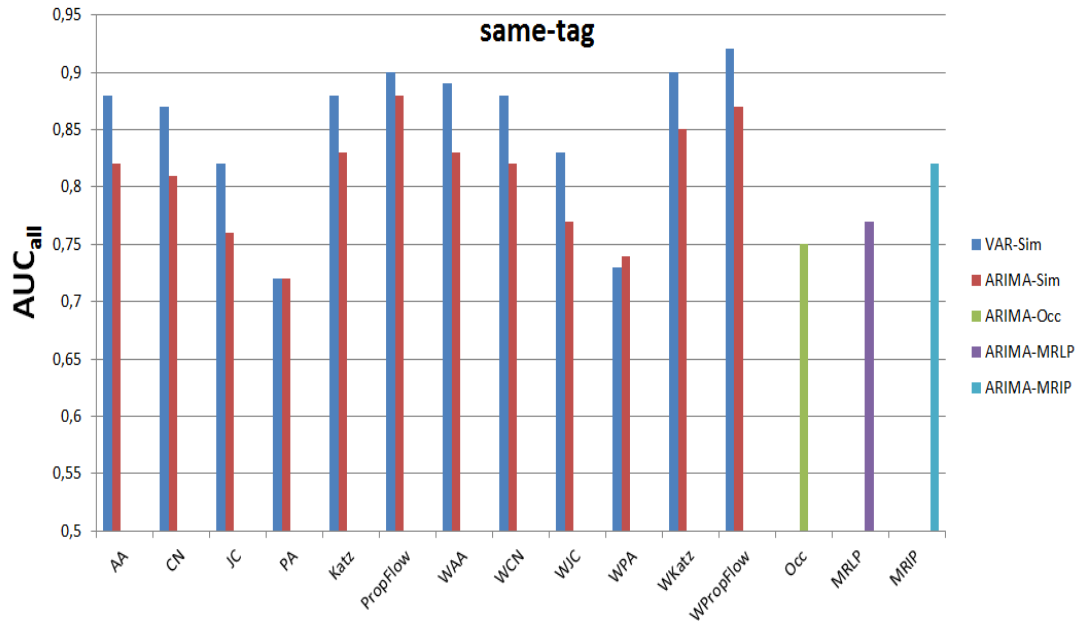
Şekil 5.6 : DBLP veri kümesi “ortak konferans” bağlantı türü için, AUC_{all} sonuçları.

Şekil 5.4, Şekil 5.5 ve Şekil 5.6’da DBLP veri kümesi üzerinde elde edilen sonuçlara göre bağlantı tahmini bağlamında, heterojen ağı oluşturan düğümler arasında meydana gelen farklı bağlantı türleri arasındaki ilişki bilgisinin ve bağlantı bilgisi ile benzerlik ölçütlerinin kombinasyonunun zaman içinde değişiminin birlikte kullanılması, düğüm benzerliklerinin ve bağlantı bilgisinin zaman içinde değişimlerinin ayrı ayrı kullanılmasından daha belirleyici olduğu ileri sürülebilir. Farklı bağlantı türleri arasındaki ilişki bilgisini kullanmak ve bağlantı bilgisi ile benzerlik ölçütleri arasındaki korelasyon bilgisini kullanmak bağlantı tahmini başarımına katkıda bulunmaktadır. Önerilen yöntem ile en iyi sonuçlar ağırlıklı PropFlow, Katz ve AA benzerlik ölçütlerinin kullanıldığı durumda elde edilirken, en düşük performans PA ölçütü ile sağlanmıştır. PA ölçütünde elde edilen düşük başarımın olası nedenleri arasında ölçütün düğümlerin komşuluk bilgisini ele almaması sadece düğümlerin derece çarpımını dikkate alması gösterilebilir. Bu nedenle çok değişkenli zaman serileri ile bağlantı bilgisi ve düğüm derece çarpımları arasındaki korelasyon bilgisini kullanmak (VAR-Sim), sadece düğüm derece çarpımlarını tek değişkenli zaman serileri ile modellenmesine (ARIMA-Sim) göre başarım artışı sağlamamıştır. Şekil 5.4, Şekil 5.5 ve Şekil 5.6’da DBLP veri kümesi üzerinde, önerilen modellerin ağırlıklı benzerlik skorları ile kullanıldığında elde edilen sonuçlar da yer almaktadır. Ağırlıklı benzerlik skorları hesaplanırken, sadece

herhangi bir zaman diliminde düğümler arasında bağlantı olup olmaması değil, bağlantıların sıklık bilgisi de dikkate alınır. Elde edilen sonuçlara göre benzerlik skorlarının ağırlıklı versiyonlarının dikkate alınması, tüm modellerin performansında artış sağlamaktadır. Bu artışın muhtemel nedeni, verilen bir zaman diliminde tekrar eden bağlantıların da dikkate alınmasıdır.



Şekil 5.7 : Delicious veri kümesi “ortak kaynak” bağlantı türü için, AUC_{all} sonuçları.

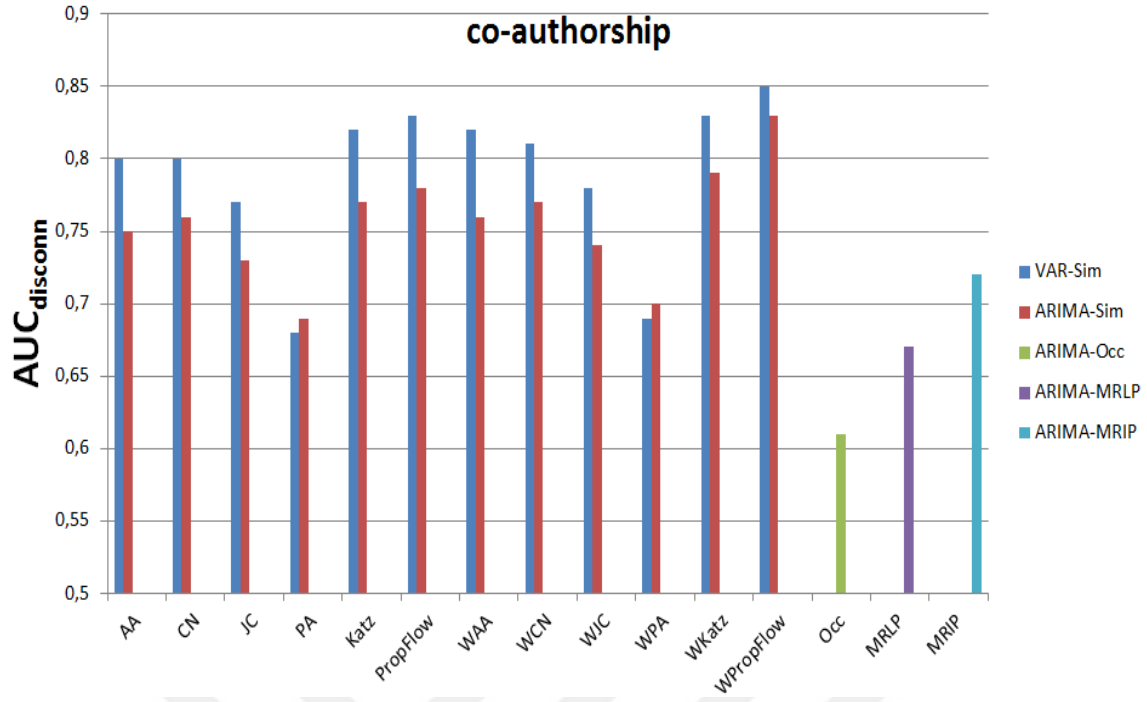


Şekil 5.8 : Delicious veri kümesi “ortak etiket” bağlantı türü için, AUC_{all} sonuçları.

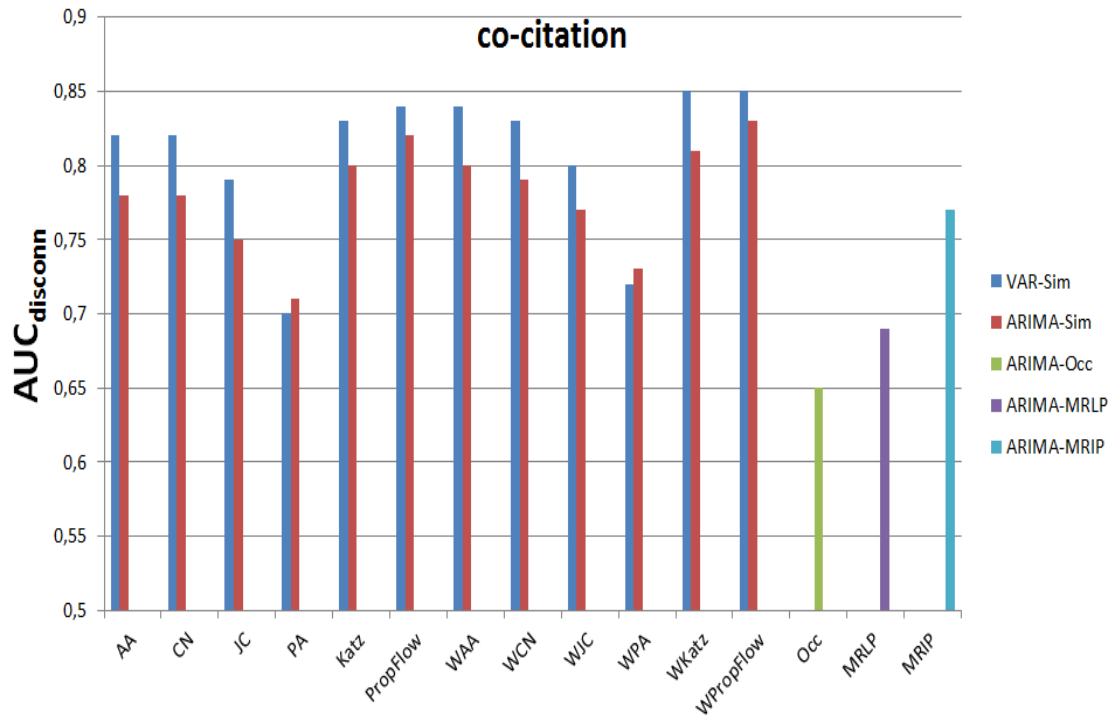
DBLP veri kümesi için yapılan testler, Delicious veri kümesi için de tekrarlanmıştır. Şekil 5.7 ve Şekil 5.8’de Delicious veri kümesi üzerinde, önceki zaman dilimlerinde bağlı ve bağlı olmayan bütün düğüm çiftleri gözetilerek hesaplanan (AUC_{all}) bağlantı tahmini sonuçları yer almaktadır. Delicious veri kümesinde kullanıcılar arasındaki “sık kullanılanlara ortak kaynakları ekleme” ve “kaynaklar için ortak etiketleri kullanma” bağlantı türlerinin temsil ettiği heterojen bir ağda bütün ilişki türlerinin tahmin edilmesi gerçekleşmiştir. Sonuçlar, kullanılan yöntemlere göre gruplanmıştır. DBLP veri kümesinde elde edilen sonuçlara benzer şekilde önerilen VAR modeli, PA dışındaki tüm benzerlik skorları için, sadece benzerlik skor bilgisinin tek değişkenli zaman serileri ile modellenmesinden ve yalnız bağlantı sıklık bilgisinin tek değişkenli zaman serileri ile modellenmesinden daha başarılı sonuç vermektedir.

Oluşturulan heterojen ağlarda farklı bağlantı türleri farklı sıklık özellikleri göstermektedir. Elde edilen sonuçlara göre bibliyografik heterojen ağdaki “ortak konferans” bağlantı türü için elde edilen başarımlar “ortak yayın” ve “ortak atıf” bağlantı türlerine göre düşüş göstermektedir benzer şekilde Delicious veri kümesi için oluşturulan heterojen ağda da “ortak kaynak” bağlantı türü için elde edilen başarımlar “ortak etiket” bağlantı türüne göre düşüş göstermektedir. Bunun olası nedeni olarak, “ortak konferans” ve “ortak kaynak” bağlantı türüne ait komşuluk bilgilerinin, ilgili veri kümelerindeki diğer bağlantı türlerine ait komşuluk bilgilerine göre seyrek yapıya sahip olması gösterilebilir. Seyrek veri kümelerinde de VAR-Sim modeli ARIMA-Sim ve ARIMA-Occ modeline göre daha iyi sonuçlar vermiştir.

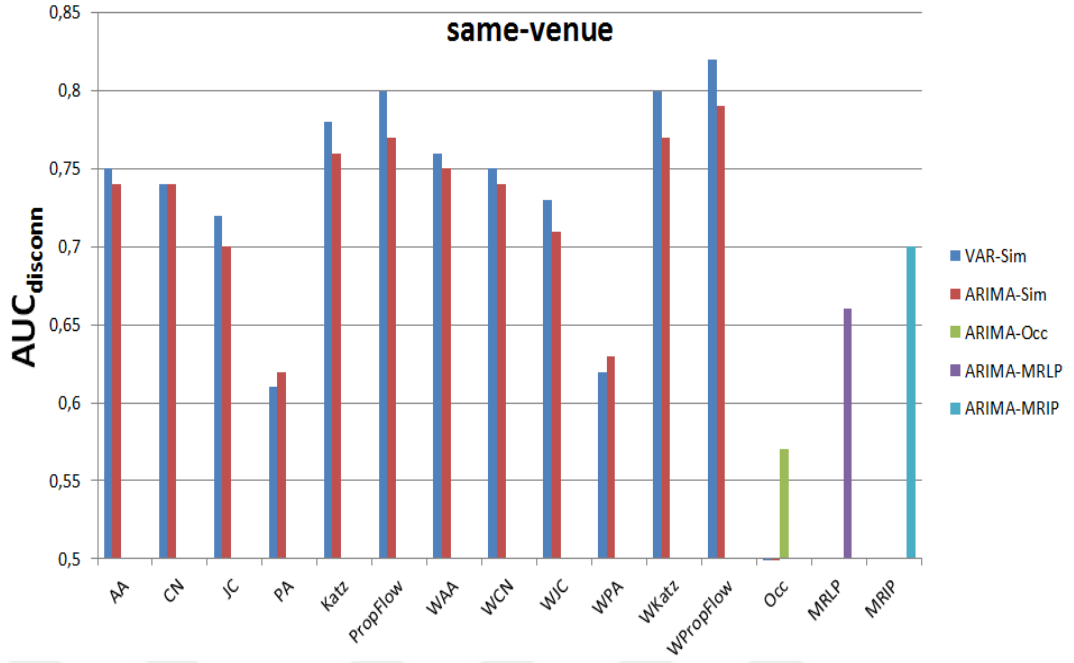
Her iki veri kümesinde elde edilen sonuçlara göre en iyi sonuçlar önerilen VAR-Sim modelinin global ölçütler olan PropFlow ve Katz ile kullanıldığı durumda elde edilmiştir. Bunun nedeni olarak, global ölçütlerin düğümler arasındaki yollar dikkate alması, sosyal ağ yapısında daha fazla topolojik bilgi elde etmesi ve tahmin etme başarımının lokal ölçütlere göre daha iyi olması gösterilebilir. Lokal ölçütler arasında ise en iyi sonuç AA ölçütünde elde edilmiştir. Özetle AUC_{all} bağlantı tahmini sonuçlarına göre önerilen VAR-Sim modeli PropFlow, Katz ve AA ölçütleri ile kullanıldığı durumda karşılaştırılan model olan ARIMA-Sim modeline göre %7’ye kadar iyileşme, ARIMA-Occ modeline göre ise %10’dan fazla iyileşme sağlamıştır.



Şekil 5.9 : DBLP veri kümesi “ortak yayın” bağlantı türü için, $AUC_{disconn}$ sonuçları.

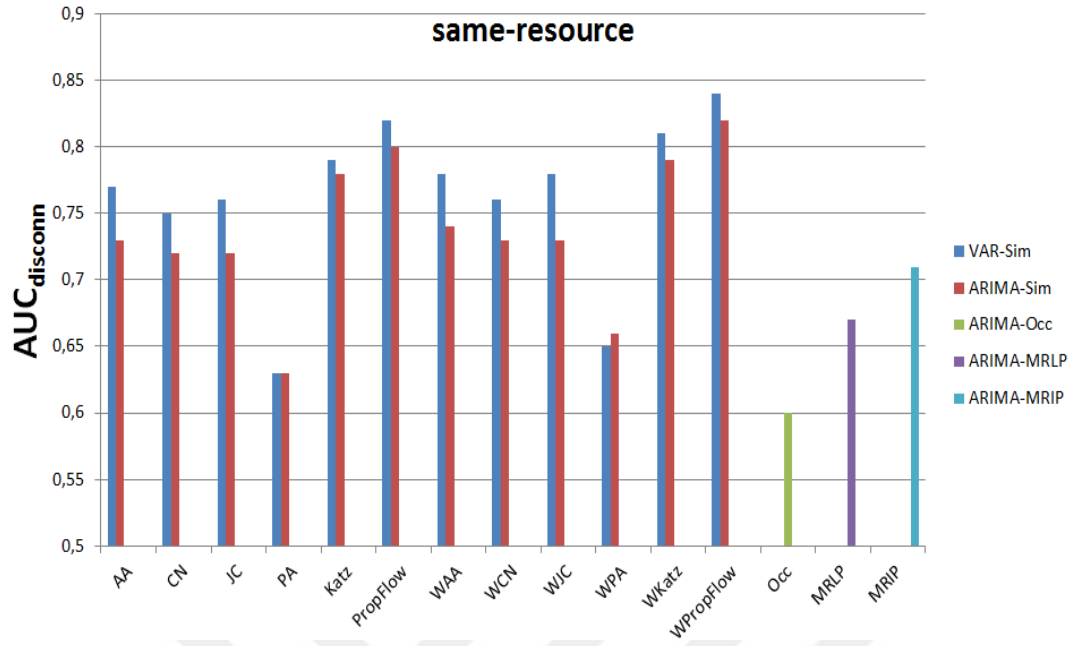


Şekil 5.10 : DBLP veri kümesi “ortak atıf” bağlantı türü için, $AUC_{disconn}$ sonuçları.

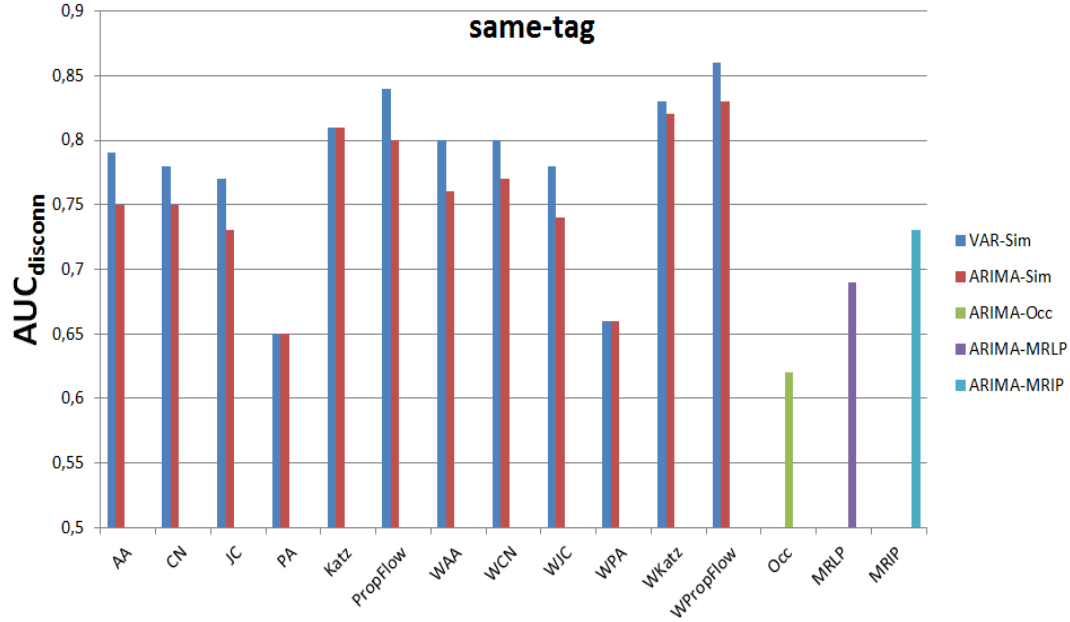


Şekil 5.11 : DBLP veri kümesi “ortak konferans” bağlantı türü için, $AUC_{disconn}$ sonuçları.

Bir düğüm çifti yakın zaman dilimlerinde bağlantı gerçekleştirilmemiş ise, ilgili düğüm çiftinin gelecekte bağlantı kurup kurmayacağını belirlemek zorlaşmaktadır. Bu durumun nedeni olarak yakın geçmiş etkisinin kullanılamaması gösterilebilir. Yakın zaman dilimlerinde bağlanmamış düğümlere ait başarıyı değerlendirmek için, son 3 zaman diliminde bağlı olmayan düğüm çiftleri ele alınmış ve bu ölçüm $AUC_{disconn}$ olarak adlandırılmıştır. Şekil 5.9, Şekil 5.10 ve Şekil 5.11’de DBLP veri kümesi üzerinde, son 3 zaman diliminde bağlı olmayan düğüm çiftleri için hesaplanan ($AUC_{disconn}$) bağlantı tahmini sonuçları yer almaktadır. Şekillerde elde edilen sonuçlara göre, önerilen çok değişkenli zaman serisi tabanlı yöntemin tek değişkenli zaman serisi tabanlı modellere göre üstünlüğü, son 3 zaman diliminde bağlı olmayan düğümler için ön plana çıkmaktadır. Bu üstünlüğün muhtemel sebebi, düğüm çiftleri arasında yakın zamanda bağlantı gerçekleşmese dahi, düğüm benzerlikleri aracılığı ile, düğüm çiftlerinin gelecekte bağlanıp bağlanmayacağı hakkında bilgi elde edilebilmesidir. Düğüm çiftleri arasındaki bağıllık bilgisini tek değişkenli zaman serilerinde kullanan (ARIMA-Occ) model ile, eğer düğüm çifti yakın geçmişte bağlantı gerçekleştirilmemiş ise ilgili düğüm çifti için bağlantı tahmini yapmak oldukça zorlaşmaktadır. Önerilen VAR-Sim modelinin PropFlow ve Katz benzerlik ölçütleri ile kullanıldığı durumda “ortak konferans” gibi seyrek yapıya sahip bağlantı türünde de %3 iyileşme sağlamıştır (Şekil 5.11).



Şekil 5.12 : Delicious veri kümesi “ortak kaynak” bağlantı türü için, $AUC_{disconn}$ sonuçları.



Şekil 5.13 : Delicious veri kümesi “ortak etiket” bağlantı türü için, $AUC_{disconn}$ sonuçları.

DBLP veri kümesi için yapılan testler, Delicious veri kümesi için de tekrarlanmıştır. Şekil 5.12 ve Şekil 5.13'te Delicious veri kümesi üzerinde, son 3 zaman diliminde bağlı olmayan düğüm çiftleri için hesaplanan ($AUC_{disconn}$) bağlantı tahmini sonuçları yer almaktadır.

Buna göre, bağlantı tahmini bağlamında, düğüm benzerliklerinin ve bağlantı bilgisinin zaman içinde değişiminin birlikte ele alınması, düğüm benzerliklerinin zaman içinde değişiminden daha belirleyici olduğu iddia edilebilir. Benzerlik ve bağlantı sıklık bilgisinin değişiminin eşanlı modellenmesi bağlantı tahmini performansına katkı sağlamaktadır. Geliştirilen model ile zaman bilgisini kullanmayan statik yöntemlerin (AA, JC, PA, CN) en büyük problemi olan zamanla değişim bilgisini kullanamama ve tekrarlı bağlantıların tahmin edilememe sorunu ile zaman bilgisini kullanan tek değişkenli zaman serisi yöntemlerinin (ARIMA-Occ) daha önce örneği görülmemeyen (meydana gelmeyen) bağlantılar ile ilgili yeterince bilgi edinememe sorunlarına çözüm üretilmiştir. İlgili problemlerin her ikisine birden çözüm üretebilmek için bağlantı bilgisi ve benzerlik ölçütleri arasında ortaya çıkan zamansal korelasyon bilgisi kullanılmış ve mevcut yöntemlerin başarımları artırılmıştır.

5.3.2.1 Farklı bağlılık koşulları altında bağlantı tahmini

Bağlantı tahmini modellerinin performansı, düğüm çiftlerinin yakın geçmişteki bağlılık durumlarına göre değişiklik gösterebilmektedir. Örneğin herhangi bir düğüm çifti yakın geçmişte birçok zaman diliminde bağlı ise ilgili düğüm çiftinin gelecekte bağlanıp bağlanmayacağını tahmin etmek kolaylaşmaktadır. Buna karşın, yakın geçmişte uzun süredir bağlılık bilgisi içermeyen düğüm çiftlerini veya bağlılık durumları sıklıkla değişen düğüm çiftleri için gelecekte bağlantı bilgisini tahmin etmek zorlaşmaktadır. Yakın geçmişte gerçekleşen farklı bağlılık koşulları altında bağlantı tahmini başarımlarını ölçebilmek için farklı AUC ölçümleri tanımlanmıştır (AUC_{all} , $AUC_{disconn}$, AUC_{conn} , $AUC_{lastDisconn}$, $AUC_{lastConn}$, AUC_{change}). Çizelge 5.7 ve Çizelge 5.8'de, önerilen VAR-Sim ve karşılaştırma amacı ile kullanılan ARIMA-Sim yönteminin, farklı AUC ölçümleri türünden sonuçları verilmiştir. Sonuçlar, kullanılan modellere göre gruplanmıştır. AUC_{all} ve $AUC_{disconn}$ sonuçlarına göre karşılaştırılan modeller arasında en iyi sonucu veren ARIMA-Sim modeli olduğu için diğer ölçümlerde VAR-Sim ve ARIMA-Sim modelleri

kullanılmıştır. Bu tablolara göre, önerilen VAR-Sim modeli, tüm benzerlik skorları için, sadece benzerlik skor bilgisinin tek değişkenli zaman serileri ile modellenmesinden daha iyi sonuç vermektedir. Önerilen yöntemler arasında, en iyi sonuçlar AA, Katz, PropFlow benzerlik ölçütleri ile elde edildiği için sadece bu benzerlikleri kullanan yöntemlerin sonuçları verilmiştir. AA, Katz, PropFlow benzerlik ölçütlerinin kullanıldığı VAR-Sim modeli ARIMA-Sim modeline göre farklı AUC ölçümlerine göre %2-11 arasında iyileştirme sağlamıştır. Önerilen yöntemlerin, farklı koşullar altında da kullanılmasının uygun olduğu iddia edilebilir

Zaman bilgisinin etkin kullanımı ve benzerlik skorları ile bağlantı bilgisinin birlikte kullanımı sayesinde, önerilen yöntem daha yüksek AUC_{conn} ve $AUC_{disconn}$ değerleri verir. Önerilen VAR-Sim modelinin, sadece benzerlik skor bilgisinin tek değişkenli zaman serileri olarak modellendiği ARIMA-Sim ile yakın $AUC_{disconn}$ ve $AUC_{lastDisconn}$ sonuçlarının vermesinin nedeni ise son zaman dilimlerinde bağlantıların eksik olması ve VAR-Sim modelinin bağlantı bilgisi ile benzerlik skorlarını yakın geçmişte birlikte kullanamamasıdır. Yakın geçmişte eksik bağlantı bilgileri içeren bir düğüm çifti ile ilgili olarak VAR-Sim modelinin çok fazla bilgi elde etmesi mümkün değildir çünkü bu yöntemde bağlantı bilgisi ile benzerlik skorlarının korelasyon bilgisi kullanılmaktadır. Son zaman dilimlerinde ve son zaman diliminde bağlı olan düğüm çiftleri için (AUC_{conn} ve $AUC_{lastConn}$) de testler yürütülmüştür. Çizelge 5.7 ve Çizelge 5.8'e göre, önerilen yöntem, her iki ölçüme göre diğer yöntemlerden her iki veri kümesinde bütün bağlantı türleri için daha başarılı sonuç vermiştir. Buradaki iyileşme önerilen VAR-Sim modelinde daha fazladır nedenleri arasında yakın geçmişteki bağlantı sıklık bilgilerinin eksik olmaması ve VAR-Sim modelinde korelasyon bilgisinin kullanılabilmesi gösterilebilir. Ayrıca, son zaman dilimine göre bağlılık durumunda değişiklik gösteren düğüm çiftleri için AUC_{change} ölçüm testleri yapılmıştır. AUC_{change} ölçümü diğer AUC ölçümlerine göre tahmin edilmesi daha zor bir durumdur. Önerilen VAR-Sim modeli, AUC_{change} için diğerlerine oranla daha düşük başarımlar sağlamasına rağmen, ARIMA-Sim modeline göre %7-11 arası başarımların artışı elde edilmiştir.

Çizelge 5.7 : DBLP veri kümesi için sonuçlar.

Bağlantı Türü	Önerilen Yöntemler VAR-Sim				Karşılaştırılan Yöntemler ARIMA-Sim		
	AUC Değerleri	WAA	Wkatz	WPropFlow	WAA	WKatz	WPropFlow
ortak-yazar	AUC _{all}	0.89	0.89	0.90	0.85	0.83	0.85
	AUC _{disconn}	0.81	0.83	0.85	0.80	0.79	0.83
	AUC _{conn}	0.73	0.75	0.78	0.66	0.67	0.68
	AUC _{lastDisconn}	0.85	0.88	0.87	0.82	0.81	0.82
	AUC _{lastConn}	0.76	0.79	0.82	0.69	0.71	0.72
	AUC _{change}	0.70	0.72	0.73	0.61	0.63	0.65
ortak-atıf	AUC _{all}	0.91	0.91	0.92	0.85	0.84	0.87
	AUC _{disconn}	0.84	0.85	0.85	0.80	0.81	0.83
	AUC _{conn}	0.76	0.77	0.80	0.68	0.69	0.71
	AUC _{lastDisconn}	0.87	0.88	0.89	0.83	0.84	0.85
	AUC _{lastConn}	0.79	0.81	0.84	0.71	0.73	0.74
	AUC _{change}	0.73	0.76	0.76	0.65	0.66	0.69
ortak-konferans	AUC _{all}	0.84	0.86	0.87	0.79	0.81	0.82
	AUC _{disconn}	0.76	0.80	0.82	0.75	0.77	0.79
	AUC _{conn}	0.67	0.69	0.70	0.59	0.61	0.63
	AUC _{lastDisconn}	0.80	0.83	0.83	0.76	0.78	0.80
	AUC _{lastConn}	0.72	0.73	0.75	0.63	0.64	0.64
	AUC _{change}	0.65	0.67	0.69	0.56	0.58	0.59

Çizelge 5.8 : Delicious veri kümesi için sonuçlar.

Bağlantı Türü	Önerilen Yöntemler VAR-Sim				Karşılaştırılan Yöntemler ARIMA-Sim		
	AUC Değerleri	WAA	Wkatz	WPropFlow	WAA	WKatz	WPropFlow
ortak-kaynak	AUC _{all}	0.86	0.88	0.89	0.79	0.83	0.85
	AUC _{disconn}	0.78	0.81	0.84	0.74	0.79	0.82
	AUC _{conn}	0.80	0.83	0.86	0.75	0.80	0.81
	AUC _{lastDisconn}	0.81	0.82	0.85	0.76	0.80	0.83
	AUC _{lastConn}	0.83	0.85	0.87	0.75	0.81	0.84
	AUC _{change}	0.71	0.72	0.74	0.63	0.64	0.65
ortak-etiket	AUC _{all}	0.89	0.90	0.92	0.83	0.85	0.87
	AUC _{disconn}	0.80	0.83	0.86	0.76	0.82	0.83
	AUC _{conn}	0.82	0.85	0.87	0.77	0.83	0.84
	AUC _{lastDisconn}	0.83	0.85	0.87	0.80	0.84	0.85
	AUC _{lastConn}	0.83	0.86	0.88	0.78	0.84	0.85
	AUC _{change}	0.74	0.77	0.78	0.64	0.65	0.67

5.3.2.2 Çapraz geçerleme testleri

Önceki bölümde gerçekleştirilen testler, 10-kat çapraz geçerleme ile tekrarlanmıştır. Daha önceki çapraz geçerleme testlerine benzer olarak zaman serilerinin oluşturulduğu zaman dilimleri, 10 eşit parçaya bölünmüş ve her bir testte ilgili parçada bulunan düğüm çiftleri kullanılmıştır. 10 farklı parçadaki düğüm çiftleri için zaman serileri modelleri oluşturulmuş, her parça için farklı bağlantı tahmini ve testler gerçekleştirilmiş ve AUC değerlerinin ortalaması son test sonucu olarak verilmiştir. DBLP ve Delicious veri kümeleri için test sonuçları Çizelge 5.9 ve Çizelge 5.10'da verilmiştir. Çapraz geçerleme ile elde edilen sonuçlar, daha az öğrenme verisi ile elde edildiği için, Bölüm 5.3.2.1'de elde edilen sonuçlara göre daha düşüktür. Önerilen modelin performans düşüşü, diğer yöntemlere göre daha azdır çünkü önerilen yöntem, daha az öğrenme kümesi ile daha yüksek performans sergilemektedir. Sonuçlarda elde edilen standart sapma değeri, önerilen modelde, diğer modele göre daha düşüktür. Bu durumun nedeni, önerilen yöntemlerde bütün bağlantı türleri için hem benzerlik skorları hem de bağlantı bilgisinin zaman içinde birlikte değişiminin kullanılmasıdır. Modeller ikili karşılaştırılarak “İlişkili ölçümler için eşli t-testi (Paired Samples t-test)” yapılmıştır. Her model için eşli t sınaması için ortalama ve standart sapma değerleri hesaplanmıştır. Bu test ile ölçümlere ait ortalamalar karşılaştırılmış, testlere ilişkin ortalamalar arasındaki farkın %95 güven düzeyinde istatistiksel olarak anlamlı olduğu da görülmüştür.

Çizelge 5.9 : DBLP veri kümesi için çapraz geçirme sonuçları.

Bağlantı Türü	Önerilen Yöntemler VAR-Sim				Karşılaştırılan Yöntemler ARIMA-Sim		
	AUC Değerleri	WAA	WKatz	WPropFlow	WAA	WKatz	WPropFlow
ortak-yazar	AUC _{all}	0.82±0.02	0.84±0.01	0.85±0.01	0.76±0.03	0.77±0.03	0.79±0.03
	AUC _{disconn}	0.75±0.02	0.77±0.01	0.78±0.01	0.69±0.02	0.72±0.03	0.74±0.02
	AUC _{conn}	0.72±0.01	0.73±0.01	0.74±0.01	0.61±0.02	0.62±0.02	0.61±0.02
	AUC _{lastDisconn}	0.78±0.02	0.82±0.02	0.81±0.02	0.73±0.02	0.75±0.02	0.76±0.02
	AUC _{lastConn}	0.74±0.01	0.75±0.01	0.76±0.01	0.64±0.03	0.66±0.02	0.67±0.02
	AUC _{change}	0.65±0.01	0.67±0.01	0.69±0.01	0.59±0.03	0.61±0.03	0.62±0.03
ortak-atıf	AUC _{all}	0.85±0.02	0.86±0.01	0.88±0.01	0.78±0.03	0.79±0.03	0.80±0.03
	AUC _{disconn}	0.77±0.02	0.79±0.01	0.78±0.01	0.72±0.02	0.73±0.03	0.76±0.02
	AUC _{conn}	0.74±0.01	0.74±0.01	0.76±0.01	0.61±0.02	0.63±0.02	0.64±0.02
	AUC _{lastDisconn}	0.81±0.02	0.83±0.02	0.84±0.02	0.75±0.02	0.76±0.02	0.77±0.02
	AUC _{lastConn}	0.75±0.01	0.76±0.01	0.77±0.01	0.67±0.03	0.67±0.02	0.69±0.02
	AUC _{change}	0.69±0.01	0.70±0.01	0.71±0.01	0.61±0.03	0.63±0.03	0.64±0.03
ortak-konferans	AUC _{all}	0.78±0.02	0.81±0.02	0.82±0.02	0.73±0.03	0.73±0.02	0.75±0.02
	AUC _{disconn}	0.71±0.02	0.74±0.02	0.75±0.02	0.66±0.03	0.67±0.02	0.68±0.02
	AUC _{conn}	0.61±0.01	0.63±0.01	0.67±0.01	0.55±0.03	0.56±0.03	0.58±0.03
	AUC _{lastDisconn}	0.75±0.02	0.76±0.02	0.77±0.02	0.69±0.02	0.72±0.02	0.72±0.02
	AUC _{lastConn}	0.68±0.01	0.69±0.01	0.70±0.01	0.59±0.02	0.60±0.02	0.62±0.02
	AUC _{change}	0.61±0.01	0.64±0.01	0.66±0.01	0.52±0.02	0.53±0.02	0.54±0.02

Çizelge 5.10 : Delicious veri kümesi çapraz geçerleme sonuçları.

Bağlantı Türü	Önerilen Yöntemler VAR-Sim				Karşılaştırılan Yöntemler ARIMA-Sim		
	AUC Değerleri	WAA	WKatz	WPropFlow	WAA	WKatz	WPropFlow
ortak-kaynak	AUC _{all}	0.79±0.02	0.82±0.01	0.83±0.01	0.73±0.03	0.74±0.03	0.75±0.03
	AUC _{disconn}	0.70±0.02	0.71±0.01	0.74±0.01	0.64±0.02	0.65±0.03	0.66±0.02
	AUC _{conn}	0.75±0.01	0.76±0.01	0.76±0.01	0.66±0.02	0.67±0.02	0.69±0.02
	AUC _{lastDisconn}	0.73±0.02	0.74±0.02	0.75±0.02	0.66±0.02	0.67±0.02	0.69±0.02
	AUC _{lastConn}	0.77±0.01	0.78±0.01	0.77±0.01	0.68±0.03	0.69±0.02	0.70±0.02
	AUC _{change}	0.66±0.01	0.65±0.01	0.69±0.01	0.58±0.03	0.59±0.03	0.60±0.03
ortak-etiket	AUC _{all}	0.82±0.02	0.84±0.01	0.85±0.01	0.76±0.03	0.77±0.03	0.79±0.03
	AUC _{disconn}	0.72±0.02	0.73±0.01	0.75±0.01	0.67±0.02	0.68±0.03	0.69±0.02
	AUC _{conn}	0.78±0.01	0.79±0.01	0.80±0.01	0.69±0.02	0.71±0.02	0.72±0.02
	AUC _{lastDisconn}	0.76±0.02	0.78±0.02	0.77±0.02	0.69±0.02	0.71±0.02	0.72±0.02
	AUC _{lastConn}	0.80±0.01	0.81±0.01	0.82±0.01	0.73±0.03	0.74±0.02	0.75±0.02
	AUC _{change}	0.69±0.01	0.70±0.01	0.73±0.01	0.59±0.03	0.62±0.03	0.63±0.03

5.3.2.3 Farklı ölçütlerin birleştirilip eş zamanlı kullanılması

Önceki bölümlerdeki test sonuçlarına göre ağırlıklı PropFlow ve AA benzerlik ölçütünü ve bağlantı bilgisini kullanan VAR-Sim modelinin ARIMA-Sim ve ARIMA-Occ modellerine göre daha başarılı sonuçlar verdiği görülmüştür. Bu bölümde lokal ölçütler arasında en iyi sonucu veren ağırlıklı AA ve global ölçütler arasında en iyi sonucu veren ağırlıklı PropFlow benzerlik skorları ve bağlantı bilgisinin eş zamanlı kullanılarak oluşturulduğu VAR-AllSim modelinin başarımı ölçülmüştür. Modeli gerçeklemek için DBLP veri kümesindeki ortak yayın, ortak atıf, ortak konferans bağlantı türleri ve ilgili bağlantı türlerinin herbiri için hesaplanan PropFlow, AA benzerlik skorları ve bağlantı bilgisinden oluşan 9-değişkenli (9-denklemlili) VAR-AllSim modeli oluşturulmuştur. Bağlantı türlerine ait benzerlik ölçütlerinden (PropFlow, AA) ve bağlantı bilgisinden oluşmakta olan 9x1 vektör türü VAR-AllSim modeline girdi olarak verilmiş ve her değişkene ait skor bilgisi çıktı olarak elde edilmiştir. Ardından bağlantı tahmini aşamasında ilgili bağlantı türüne ait elde edilen ölçüt skor değerlerinin ortalaması bağlantı tahmini aşamasında kullanılmıştır. Benzer şekilde Delicious veri kümesindeki ortak kaynak, ortak atıf bağlantı türlerine ait PropFlow, AA benzerlik skorları ve bağlantı bilgisinden oluşan 6-değişkenli VAR-AllSim modeli oluşturulmuştur. 6x1 vektör türü kullanılarak benzerlik skorlarının gelecek skor değerleri tahmin edilmiş ve ilgili bağlantı türüne ait elde edilen ölçüt skor değerlerinin ortalaması bağlantı tahminin aşamasında kullanılmıştır. Çizelge 5.11 ve Çizelge 5.12’de veri kümelerine ait bütün bağlantı türleri için elde edilen farklı AUC sonuçları VAR-AllSim (PropFlow, AA) için verilmiştir.

Çizelge 5.11 : DBLP veri kümesi bütün bağlantı türleri için farklı AUC sonuçları.

Önerilen Yöntem VAR-AllSim (PropFlow, AA)						
Bağlantı türü	AUC_{all}	AUC_{disconn}	AUC_{conn}	AUC_{lastDisconn}	AUC_{lastConn}	AUC_{change}
ortak-yayın	0.91	0.85	0.79	0.88	0.83	0.73
ortak-atıf	0.93	0.85	0.81	0.88	0.85	0.77
ortak-konferans	0.88	0.82	0.71	0.83	0.75	0.68

Çizelge 5.12 : Delicious veri kümesi bütün bağlantı türleri için farklı AUC sonuçları.

Önerilen Yöntem VAR-AllSim (PropFlow, AA)						
Bağlantı türü	AUC_{all}	$AUC_{disconn}$	AUC_{conn}	$AUC_{lastDisconn}$	$AUC_{lastConn}$	AUC_{change}
ortak-kaynak	0.89	0.83	0.85	0.84	0.86	0.75
ortak-etiket	0.91	0.86	0.87	0.87	0.88	0.79

Ardından önceki bölümlerdeki testlerde en iyi sonucun elde edildiği ağırlıklı PropFlow, Katz ve AA benzerlik skorları kullanılarak elde edilen sonuçlar Çizelge 5.13 ve Çizelge 5.14’te verilmiştir. Modeli gerçeklemek için DBLP veri kümesindeki ortak yayın, ortak atıf, ortak konferans bağlantı türleri ve ilgili bağlantı türlerinin herbiri için hesaplanan PropFlow, Katz ve AA benzerlik skorları ve bağlantı bilgisinden oluşan 12-değişkenli (12-denklemlı) VAR-AllSim modeli oluşturulmuştur. Benzer şekilde Delicious veri kümesindeki ortak kaynak, ortak atıf bağlantı türlerine ait PropFlow, Katz, AA benzerlik skorları ve bağlantı bilgisinden oluşan 8-değişkenli VAR-AllSim modeli oluşturulmuştur. Elde edilen sonuçlara göre farklı bağlantı türleri arasındaki ilişki bilgisini eş zamanlı kullanmak, farklı benzerlik ölçütleriyle bağlantı bilgisinin aralarındaki korelasyon bilgisini kullanmak bağlantı tahmini başarımına katkıda bulunmaktadır.

Çizelge 5.13 : DBLP veri kümesi bütün bağlantı türleri için farklı AUC sonuçları.

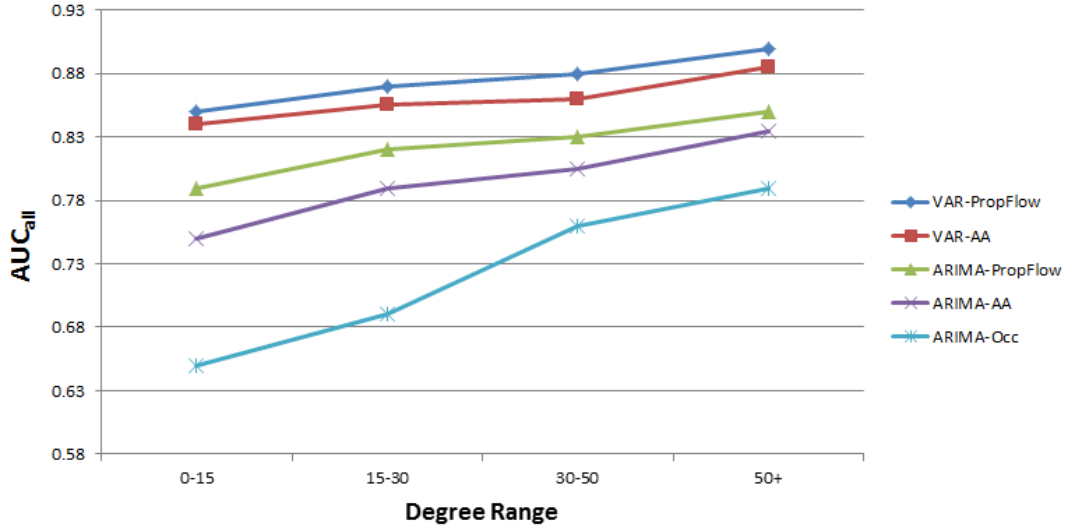
Önerilen Yöntem VAR-AllSim (PropFlow, Katz, AA)						
Bağlantı türü	AUC_{all}	$AUC_{disconn}$	AUC_{conn}	$AUC_{lastDisconn}$	$AUC_{lastConn}$	AUC_{change}
ortak-yayın	0.93	0.87	0.81	0.90	0.84	0.74
ortak-atıf	0.94	0.87	0.82	0.90	0.86	0.79
ortak-konferans	0.90	0.83	0.72	0.84	0.76	0.70

Çizelge 5.14 : Delicious veri kümesi bütün bağlantı türleri için farklı AUC sonuçları.

Önerilen Yöntem VAR-AllSim (PropFlow, Katz, AA)						
Bağlantı türü	AUC_{all}	$AUC_{disconn}$	AUC_{conn}	$AUC_{lastDisconn}$	$AUC_{lastConn}$	AUC_{change}
ortak-kaynak	0.90	0.86	0.87	0.86	0.88	0.76
ortak-etiket	0.93	0.87	0.88	0.88	0.89	0.80

5.3.2.4 Dügüm derecesinin etki analizi

Bağlantı tahmini yöntemlerinin başarımı, düğümler arasındaki geçmiş bağlantıların sıklığına bağlıdır (Lü ve Zhou, 2011). Seyrek yapıya sahip sosyal ağlarda bağlantı tahmini yoğun ağlara göre zorlaşmaktadır. Muhtemel nedenleri arasında seyrek yapıya sahip olan ağlarda düğüm çiftleri arasında benzerlik skorları hesaplanırken yeterli sayıda komşuluk bilgisinin elde edilememesi gösterilebilir. Bu sebeple ağın yapısındaki seyreklilik oranı arttıkça düğüm çiftlerinin birbirine bağlanma olasılıkları azalmaktadır. Farklı bağlantı sıklık değeri ile bağlantı tahmini başarımını değerlendirebilmek için, belirli sayıda derecesi olan düğüm kümeleri ile test yapılmaktadır. Dügüm derecelerinin az olduğu (az komşusu olan) düğüm çiftleri arasında, geçmiş bağlantıların sıklığı da azdır. Hem derece sayısı az düğümler hem de derece sayısı fazla düğümler arasında bağlantı tahmini başarımının değişimi analiz edilmiştir. Yüksek performans sağlayan bağlantı tahmini modelinin her iki durumda da başarılı sonuç göstermesi beklenir. Seyrek yapıdan yoğun yapıya doğru değişen çeşitli durumlar için testler gerçekleştirilmiştir. Ortak-yayın bağlantı türüne ait farklı derece aralıkları için modellerin başarımları Şekil 5.14'te verilmektedir. Şekildeki düğüm derecelerini gösteren her aralık, benzer sayıda düğüm sayısını içerecek şekilde tasarlanmıştır. Önerilen ve karşılaştırılan modeller, spesifik derece aralığındaki düğümler arası bağlantıların tahmin edilmesinde kullanılmıştır. Sadece bağlantı sıklık bilgisinin tek değişkenli zaman serilerinde kullanımını içeren ARIMA-Occ modelinin başarımı, derece sayısı az olan düğümler için önemli ölçüde azalmaktadır. Muhtemel sebepleri arasında, derece sayısı az olan düğümler arasında geçmiş bağlantı sıklık değerinin az olması gösterilebilir. Sadece ağırlıklı PropFlow ve AA benzerliğini kullanan ARIMA-WPropFlow ve ARIMA-WAA modellerinin başarımı ise ARIMA-Occ'ye göre daha az düşmektedir. Diğer yandan, önerilen yöntemler olan ağırlıklı PropFlow ve AA benzerliğini ve bağlantı bilgisini kullanan VAR modeli (VAR-WPropFlow ve VAR-WAA) bir miktar düşmektedir. Bu durumun olası nedenleri arasında, bağlantı sıklık değişim bilgisine ilaveten düğüm benzerlik skorlarının değişim bilgisinin de kullanılması gösterilebilir. Geliştirilen model, düğüm çiftleri arasında bağlantı olmasa dahi başarılı tahminler gerçekleştirmektedir.



Şekil 5.14 : Düğüm popülerliğinin (derecesinin) etkisi.

5.3.2.5 Karmaşıklık analizi

Geliştirilen model, bütün düğümler için komşuluk listesini tutan bir komşu düğümler listesi ve bütün düğüm çiftleri için ortak komşular listesi yaratarak hesaplama işlemine başlamaktadır. Oluşturulan listeler aracılığı ile düğüm benzerlik skorları, en az sayıda işlemle gerçekleştirilmektedir. Düğüm çiftlerinin tamamı için benzerlik skorlarının hesaplanması N düğüm sayısını göstermek üzere, $O(N^3)$ karmaşıklık ile gerçekleştirilmektedir. Gelecek benzerlik skorlarının tahmin edilme aşamasında ise, “ p ” adet gecikme değeri kullanılmaktadır. VAR modelinde skor değeri tahmin edilecek değişken sayısı (n , değişken sayısıdır) kadar denklem sayısı olduğu için VAR modelinin analiz aşamasında ele alınan değişken sayısı kadar denklem eş zamanlı çözümlenmektedir böylece gelecek skor değerlerinin tahmin edilmesi için $O(N^2(p * n^2))$ adet işleme ihtiyaç duyulmaktadır. Tahmin edilen skor değerlerinin sıralanması işleminin karmaşıklığı da $O(\log(N^2))$ olarak verilebilir. Sonuç olarak önerilen modelin karmaşıklığı, yürütülen maksimum işlem sayısı ile ifade edilmekte ve lokal benzerlik ölçütlerin kullanıldığı durumda alttaki gibi verilmektedir.

$$Karmaşıklık(N, p, n) = \begin{cases} (p * n^2) > N \text{ ise} & O(N^2(p * n^2)) \\ \text{değilse} & O(N^3) \end{cases}$$

Katz ölçütünün kullanıldığı durumda ise şu şekilde verilir:

$$Karmaşıklık(N, p, n) = \begin{cases} (p * n^2) > N \text{ ise} & O(N^2(p * n^2)) \\ \text{değilse} & O(N^3) \end{cases}$$

PropFlow ölçütünün kullanıldığı durumda ise (E , kenar sayısıdır) şu şekilde verilir:

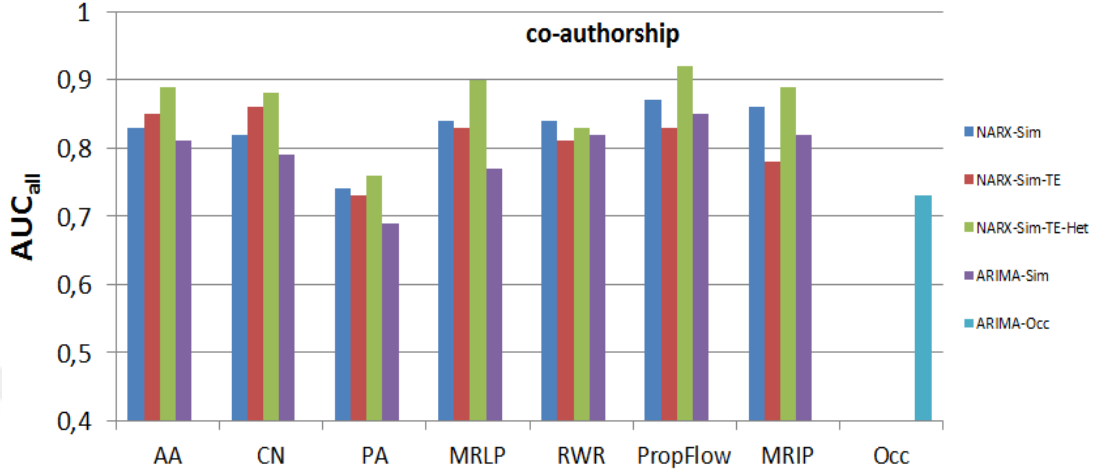
$$Karmaşıklık(N, p, n) = \begin{cases} (p * n^2) > N + E \text{ ise} & O(N^2(p * n^2)) \\ \text{değilse} & O(N^2(N + E)) \end{cases}$$

5.3.3 NARX modeli ile heterojen ağlarda bağlantı tahmini testleri

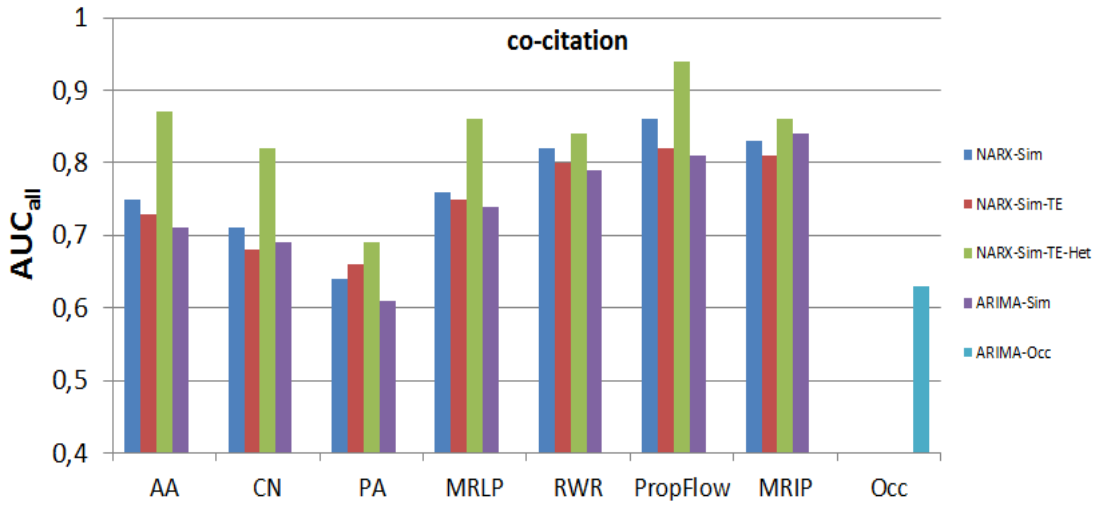
Bu bölümdeki bağlantı tahmini testlerinin temel amacı, önerilen NARX modelleri (NARX-Sim, NARX-Sim-TE, NARX-Sim-TE-Het) ile mevcut tek değişkenli zaman serileri modellerinin (ARIMA-Sim ve ARIMA-Occ) bağlantı tahmini performanslarını karşılaştırmaktır. “Sim” zaman serileri modellerinde olduğu gibi benzerlik ölçütünü, “TE” ise bağlantı sıklık bilgisindeki değişimleri (oluşan, korunan, kopan) göstermektedir. NARX-Sim modelinde ilgili bağlantı türü için benzerlik skoru girdi olarak kullanılırken, NARX-Sim-TE’de ilgili bağlantı türü için benzerlik skorları ve bağlantı sıklık bilgisindeki değişimler girdi olarak kullanılmıştır. NARX-Sim-TE-Het modelinde ise bütün bağlantı türleri için hesaplanan benzerlik skorları ve bağlantı sıklık bilgisindeki değişimler girdi olarak kullanılmıştır. Bu yöntemlerde, tüm benzerlik ölçütleri için ağırlıklı versiyonlar kullanılmıştır. Çalışma kapsamında ağ görüntüleri arasında düğümlerin değişmediği sadece bağlantı bilgisinin değiştiği kabul edilmiştir.

Şekil 5.15, Şekil 5.16 ve Şekil 5.17’de DBLP veri kümesi üzerinde, önceki zaman dilimlerinde bağlı ve bağlı olmayan bütün düğüm çiftleri gözetilerek hesaplanan (AUC_{all}) bağlantı tahmini sonuçları yer almaktadır. Sonuçlar, kullanılan yöntemlere göre gruplanmıştır. Bu tablolara göre, önerilen NARX-Sim-TE-Het modeli, tüm benzerlik skorları için, sadece benzerlik skor bilgisinin tek değişkenli zaman serileri ile modellenmesinden ve yalnız bağlantı sıklık bilgisinin tek değişkenli zaman serileri ile modellenmesinden daha başarılı sonuç vermektedir. NARX-Sim ve NARX-Sim-TE modelleri karşılaştırıldığında ise NARX-Sim modelinin daha başarılı olduğu görülmüştür. Nedenleri arasında NARX-Sim-TE “kopan bağlantı sayısı” bilgisinden olumsuz etkilenmesi gösterilebilir. İlgili bağlantı türleri için NARX modeline girdi olarak verilen değişkenlerin sonuca katkıları bağlantı ağırlık algoritmasına göre (connection weight algorithm) belirlenmiştir. Giriş olarak verilen

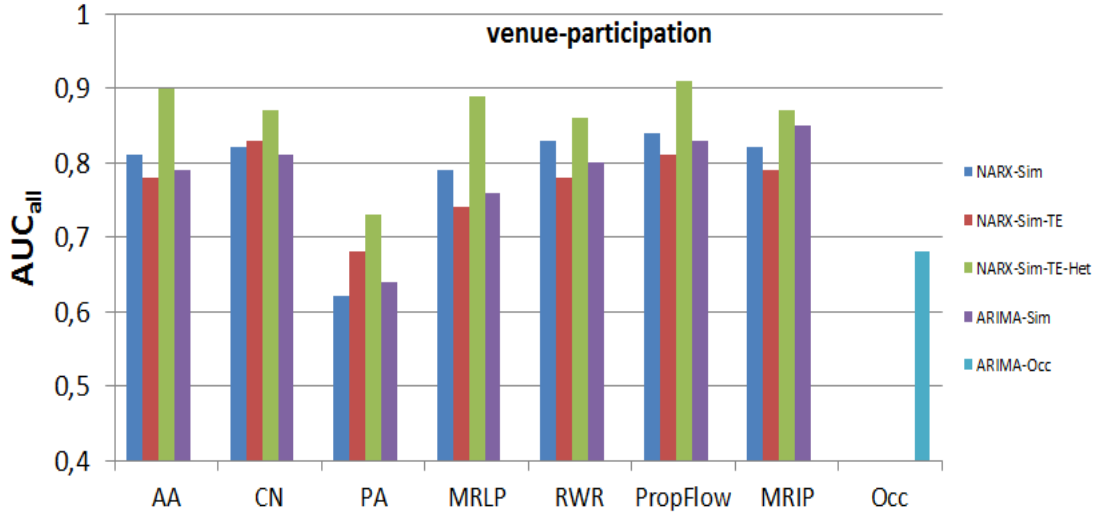
değişkenlerin katkıları incelendiğinde ilgili bağlantı türleri için “kopan bağlantı sayısı” bilgisi en az katkı sağlayan değişken olarak belirlenmiştir. Bölüm 5.3.3.3’te göreceli önem dereceleri analizinde giriş değişkenlerinin katkılarının detayları verilmiştir.



Şekil 5.15 : DBLP veri kümesi “ortak yayın” bağlantı türü için, AUC_{all} sonuçları.

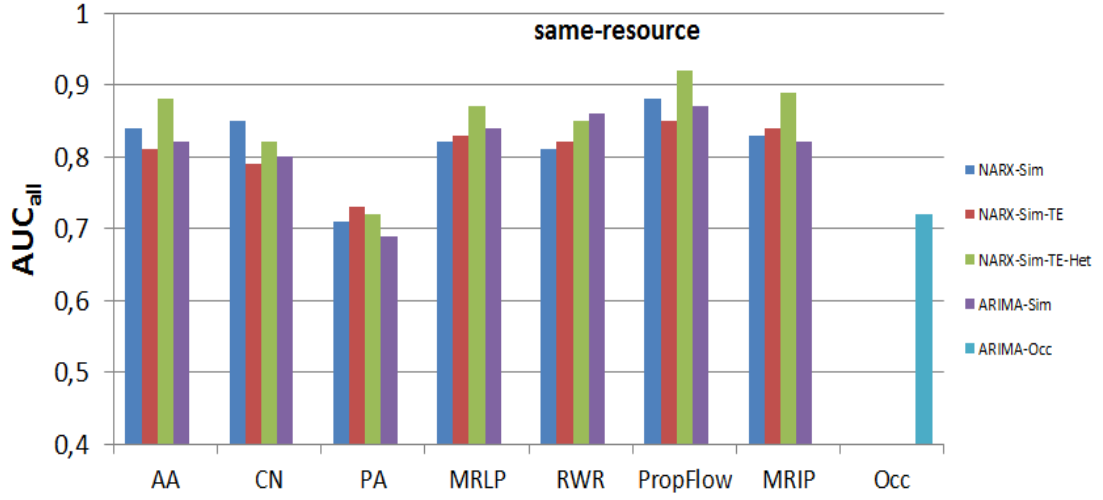


Şekil 5.16 : DBLP veri kümesi “ortak atıf” bağlantı türü için, AUC_{all} sonuçları.

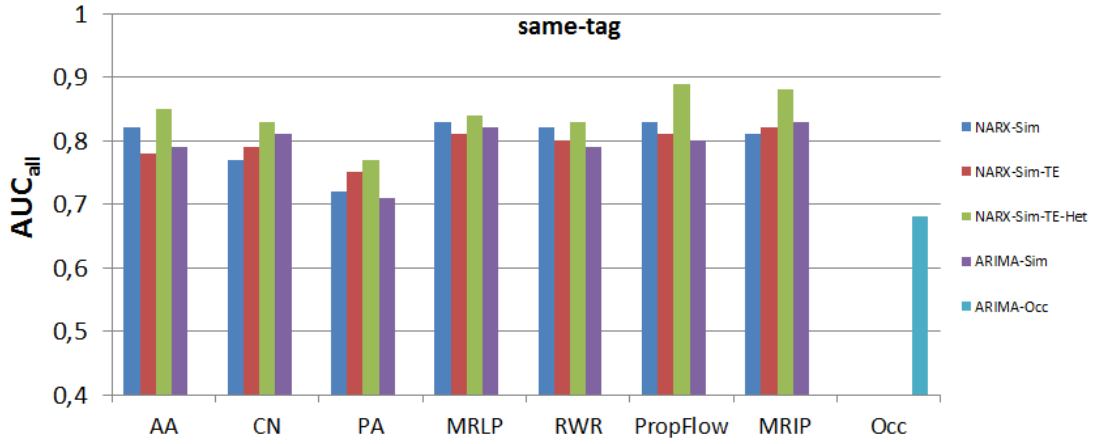


Şekil 5.17 : DBLP veri kümesi “ortak konferans” bağlantı türü için, AUC_{all} sonuçları.

Şekil 5.15, Şekil 5.16 ve Şekil 5.17’de DBLP veri kümesi üzerinde benzerlik ölçütlerinin ağırlıklı versiyonlarının kullanıldığı durumda elde edilen sonuçlar verilmiştir. Ağırlıklı skorların hesaplanması aşamasında, sadece düğüm çiftleri arasında herhangi bir zaman diliminde bağlantı gerçekleşip gerçekleşmemesi (var/yok) yerine, bağlantı sıklık bilgisi dikkate alınmaktadır. Elde edilen sonuçlara göre bağlantı tahmini bağlamında, heterojen ağı oluşturan düğümler arasında meydana gelen farklı bağlantı türleri arasındaki ilişki bilgisinin ve bağlantı sıklık değişim bilgisi ile benzerlik ölçütlerinin kombinasyonunun zaman içinde değişiminin birlikte kullanılması, sadece düğüm benzerliklerinin zaman içinde değişiminden ve yalnız bağlantı sıklık bilgisinin zaman içinde değişiminden daha belirleyici olduğu ileri sürülebilir. Farklı bağlantı türleri arasındaki ilişki bilgisini kullanmak ve bağlantı bilgisi ile benzerlik ölçütleri arasındaki korelasyon bilgisini kullanmak bağlantı tahmini başarımına katkıda bulunmaktadır. Önerilen yöntemde en başarılı sonuç ağırlıklı PropFlow, MRIP, MRLP ve AA benzerlik ölçütlerinin kullanıldığı durumda elde edilirken, en düşük performans PA ölçütü ile elde edilmektedir. PA ölçütünde elde edilen düşük başarımın olası nedenleri arasında ölçütün düğümlerin komşuluk bilgisini ele almaması sadece düğümlerin derece çarpımını dikkate alması gösterilebilir.



Şekil 5.18 : Delicious veri kümesi “ortak kaynak” bağlantı türü için, AUC_{all} sonuçları.



Şekil 5.19 : Delicious veri kümesi “ortak etiket” bağlantı türü için, AUC_{all} sonuçları.

DBLP veri kümesi için yapılan testler, Delicious veri kümesi için de tekrarlanmıştır. Şekil 5.18 ve Şekil 5.19’da Delicious veri kümesi üzerinde, önceki zaman dilimlerinde bağlı ve bağlı olmayan bütün düğüm çiftleri gözetilerek hesaplanan (AUC_{all}) bağlantı tahmini sonuçları yer almaktadır. Delicious veri kümesinde kullanıcılar arasındaki “sık kullanılanlara ortak kaynakları ekleme” ve “kaynaklar için ortak etiketleri kullanma” bağlantı türlerinin temsil ettiği heterojen bir ağda bütün ilişki türlerinin tahmin edilmesi gerçekleştirilmiştir. Sonuçlar, kullanılan yöntemlere göre gruplanmıştır. DBLP veri kümesinde elde edilen sonuçlara benzer şekilde önerilen NARX-Sim-TE-Het modeli, tüm benzerlik skorları için, sadece benzerlik skor bilgisinin tek değişkenli zaman serileri ile modellenmesinden ve yalnız bağlantı sıklık bilgisinin tek değişkenli zaman serileri ile modellenmesinden daha başarılı sonuç vermektedir.

Her iki veri kümesinde elde edilen sonuçlara göre en iyi sonuçlar önerilen NARX-Sim-TE-Het modelinin global ölçütler olan PropFlow ve MRIP ile kullanıldığı durumda elde edilmiştir. Bunun olası nedeni olarak, global ölçütlerin düğümler arasındaki yollar dikkate alması, sosyal ağ yapısında daha fazla topolojik bilgi elde etmesi ve tahmin etme başarımının lokal ölçütlere göre daha iyi olması gösterilebilir. Lokal ölçütler arasında ise en iyi sonuç AA ve MRLP metriğinde elde edilmiştir. Özetle **AUC_{all}** bağlantı tahmini sonuçlarına göre önerilen NARX-Sim-TE-Het modeli PropFlow, MRIP, AA ve MRLP ölçütleri ile kullanıldığı durumda karşılaştırılan model olan ARIMA-Sim modeline göre %15'e kadar iyileşme, ARIMA-Occ modeline göre ise %20'den fazla iyileşme sağlamıştır.

5.3.3.1 Farklı bağlılık koşulları altında bağlantı tahmini

Çizelge 5.15 ve Çizelge 5.16'da, geliştirilen ve karşılaştırma amacı ile kullanılan bazı modellerin, değişik AUC ölçümleri açısından elde edilen sonuçları görülmektedir. Sonuçlar, kullanılan yöntemlere göre gruplanmıştır. **AUC_{all}** sonuçlarına göre karşılaştırılan modeller arasında en iyi sonucu veren ARIMA-Sim modeli olduğu için diğer ölçümlerde NARX-Sim-TE-Het ve ARIMA-Sim modelleri kullanılmıştır. Bu tabloya göre, önerilen NARX-Sim-TE-Het modeli, tüm benzerlik skorları için, sadece benzerlik skor bilgisinin tek değişkenli zaman serileri ile modellenmesinden daha iyi sonuç vermektedir. Önerilen yöntemler arasında, en iyi sonuçlar AA, MRLP, PropFlow ve MRIP benzerlik ölçütleri ile elde edildiği için sadece bu benzerlikleri kullanan yöntemlerin sonuçları verilmiştir. AA, MRLP, PropFlow ve MRIP benzerlik ölçütlerinin kullanıldığı NARX-Sim-TE-Het modeli ARIMA-Sim modeline göre farklı AUC ölçümlerine göre; %5'ten fazla **AUC_{disconn}**, %10'dan fazla **AUC_{conn}**, %12'den fazla **AUC_{change}** ölçüm türünde iyileştirme sağlamıştır. Önerilen yöntemlerin, farklı koşullar altında da kullanılmasının uygun olduğu iddia edilebilir.

Çizelge 5.15 : DBLP veri kümesi bütün bağlantı türleri için farklı AUC sonuçları.

Bağlantı Türü	Önerilen Yöntemler NARX-Sim-TE-Het					Karşılaştırılan Yöntemler ARIMA-Sim			
	AUC Değerleri	AA	MRLP	PropFlow	MRIP	AA	MRLP	PropFlow	MRIP
ortak-yazar	AUC _{all}	0.89±0.02	0.90±0.01	0.92±0.01	0.89±0.01	0.81±0.03	0.77±0.02	0.85±0.02	0.82±0.03
	AUC _{disconn}	0.78±0.01	0.76±0.01	0.84±0.01	0.85±0.01	0.71±0.02	0.73±0.02	0.76±0.02	0.74±0.03
	AUC _{conn}	0.82±0.01	0.80±0.01	0.87±0.01	0.84±0.01	0.65±0.03	0.64±0.03	0.66±0.02	0.65±0.03
	AUC _{change}	0.75±0.01	0.74±0.01	0.81±0.01	0.78±0.01	0.61±0.03	0.62±0.02	0.63±0.02	0.64±0.03
ortak-atıf	AUC _{all}	0.87±0.01	0.86±0.01	0.94±0.01	0.86±0.01	0.71±0.02	0.74±0.03	0.81±0.03	0.84±0.03
	AUC _{disconn}	0.74±0.02	0.72±0.01	0.84±0.02	0.83±0.02	0.65±0.04	0.69±0.03	0.75±0.03	0.74±0.03
	AUC _{conn}	0.78±0.01	0.73±0.01	0.89±0.01	0.85±0.01	0.61±0.02	0.64±0.03	0.71±0.02	0.72±0.02
	AUC _{change}	0.79±0.01	0.75±0.01	0.78±0.01	0.77±0.01	0.55±0.02	0.58±0.03	0.64±0.02	0.63±0.03
ortak-konferans	AUC _{all}	0.90±0.01	0.89±0.01	0.91±0.01	0.87±0.01	0.79±0.02	0.76±0.03	0.83±0.02	0.85±0.02
	AUC _{disconn}	0.76±0.02	0.75±0.01	0.81±0.02	0.79±0.02	0.66±0.03	0.67±0.03	0.73±0.03	0.74±0.03
	AUC _{conn}	0.79±0.01	0.76±0.01	0.85±0.01	0.86±0.01	0.63±0.03	0.62±0.03	0.70±0.03	0.68±0.03
	AUC _{change}	0.77±0.01	0.78±0.01	0.83±0.01	0.81±0.01	0.59±0.02	0.57±0.02	0.61±0.02	0.64±0.02

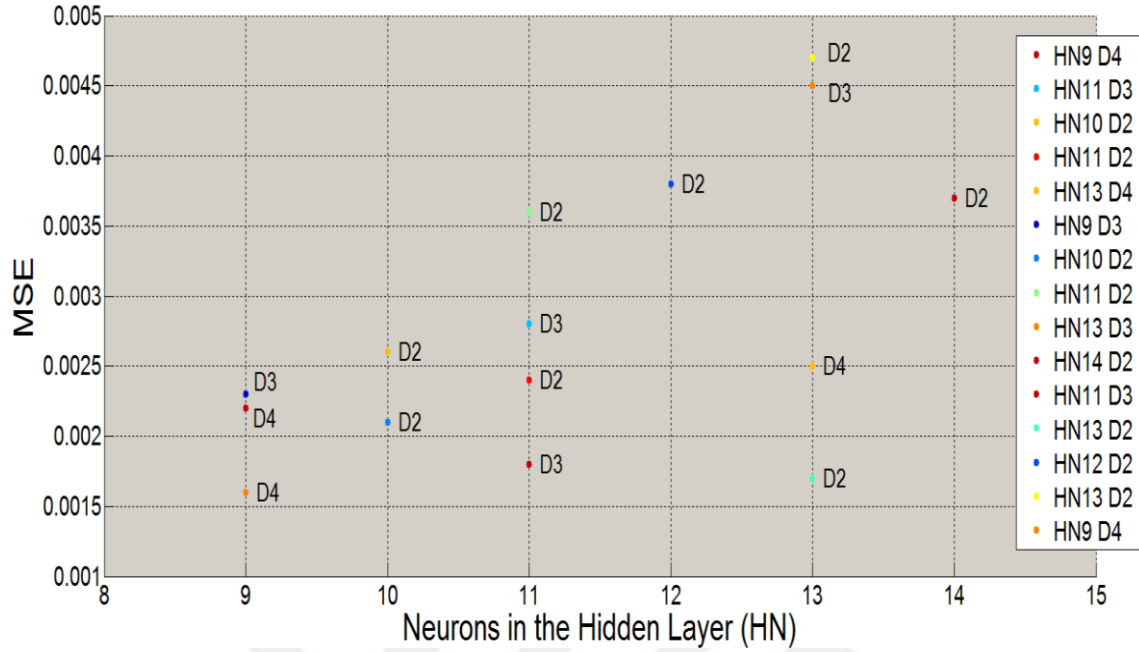
Çizelge 5.16 : Delicious veri kümesi bütün bağlantı türleri için farklı AUC sonuçları.

Bağlantı Türü	Önerilen Yöntemler NARX-Sim-TE-Het					Karşılaştırılan Yöntemler ARIMA-Sim			
	AUC Değerleri	AA	MRLP	PropFlow	MRIP	AA	MRLP	PropFlow	MRIP
ortak-yayın	AUC _{all}	0.88±0.01	0.87±0.01	0.92±0.01	0.89±0.01	0.82±0.02	0.84±0.02	0.87±0.02	0.82±0.02
	AUC _{disconn}	0.76±0.01	0.73±0.01	0.85±0.01	0.82±0.01	0.73±0.02	0.74±0.03	0.78±0.03	0.77±0.03
	AUC _{conn}	0.81±0.01	0.79±0.01	0.86±0.01	0.88±0.01	0.69±0.03	0.67±0.03	0.72±0.02	0.73±0.02
	AUC _{change}	0.77±0.01	0.76±0.02	0.80±0.01	0.78±0.01	0.57±0.03	0.61±0.02	0.65±0.02	0.66±0.03
ortak-etiket	AUC _{all}	0.85±0.01	0.84±0.01	0.89±0.01	0.88±0.02	0.79±0.03	0.82±0.02	0.80±0.03	0.83±0.03
	AUC _{disconn}	0.72±0.01	0.74±0.01	0.80±0.02	0.81±0.01	0.69±0.02	0.70±0.03	0.75±0.02	0.73±0.03
	AUC _{conn}	0.77±0.01	0.75±0.01	0.84±0.01	0.82±0.01	0.64±0.03	0.62±0.03	0.71±0.02	0.69±0.03
	AUC _{change}	0.71±0.01	0.73±0.01	0.81±0.01	0.76±0.01	0.54±0.03	0.56±0.03	0.62±0.02	0.58±0.03

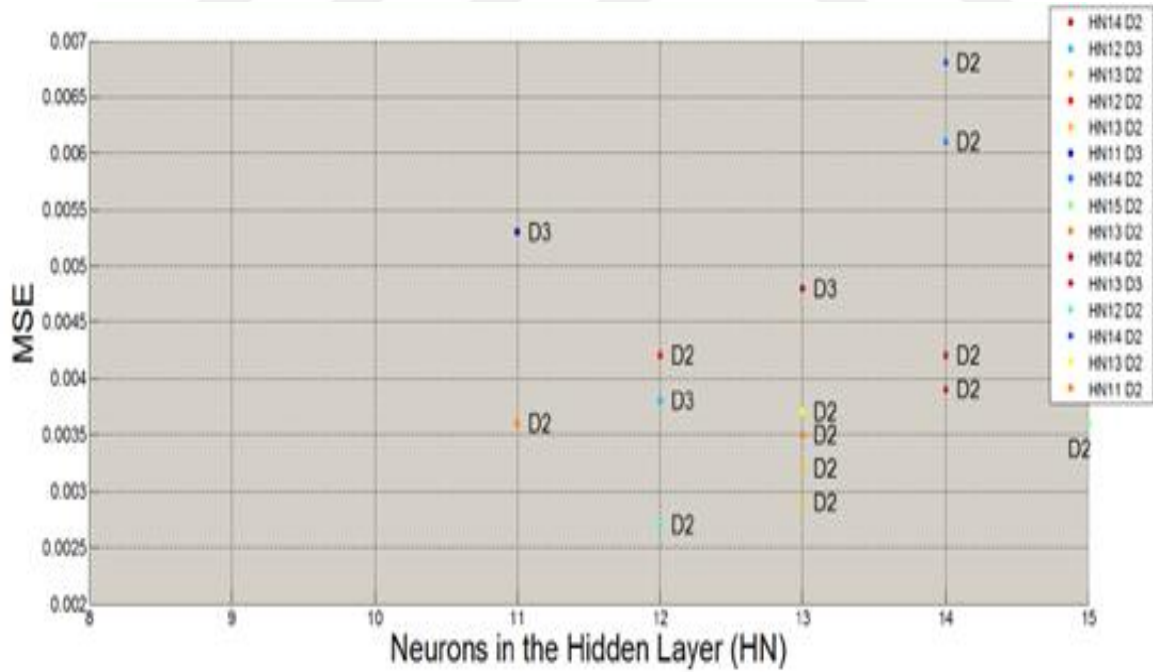
Zaman bilgisinin etkin kullanımı ve benzerlik skorları ile bağlantı bilgisinin birlikte kullanımı sayesinde, önerilen yöntem daha yüksek AUC_{conn} ve $AUC_{disconn}$ değerleri verirler. Çizelge 5.15 ve Çizelge 5.16'ya göre, önerilen model, her iki ölçüme göre diğer yöntemlerden her iki veri kümesinde bütün bağlantı türleri için daha iyi sonuç vermiştir. AUC_{conn} ölçüm türündeki iyileşme önerilen NARX-Sim-TE-Het modelinde daha fazladır muhtemel nedenleri arasında yakın geçmişteki bağlantı bilgilerinin eksik olmaması ve NARX-Sim-TE-Het modelinde korelasyon bilgisinin kullanılabilmesi gösterilebilir. Ayrıca, son zaman dilimine göre bağlılık durumunda değişiklik gösteren düğüm çiftleri için (AUC_{change}) de testler yapılmıştır. Önerilen NARX-Sim-TE-Het modeli, AUC_{change} için diğerlerine oranla daha düşük başarımla sağlanmasına rağmen, ARIMA-Sim modeline göre %12'den fazla başarımlı artış elde edilmiştir. Test sonuçlarının değerlendirilmesi için modeller ikili karşılaştırılarak “İlişkili ölçümler için eşli t -testi (Paired Samples t -test)” yapılmıştır. Her model için eşli t sınaması için ortalama ve standart sapma değerleri hesaplanmıştır. Bu test ile ölçümlere ait ortalamalar karşılaştırılmış, testlere ilişkin ortalamalar arasındaki farkın %95 güven düzeyinde istatistiksel olarak anlamlı olduğu da görülmüştür.

Testlerde, 10-kat çapraz geçerleme kullanılmıştır. Zaman serilerinin oluşturulduğu (öğrenme kümesi) son zaman dilimleri, 10 eşit parçaya ayrılmış ve her bir testte ilgili parçada bulunan düğüm çiftleri kullanılmıştır. Böylece 10 farklı parçadaki düğüm çiftleri için zaman serileri modelleri oluşturulmuş ve her parça için farklı bağlantı tahmini ve testler gerçekleştirilmiştir. Bu 10 farklı kümedeki testler sonucunda elde edilen AUC değerlerinin ortalaması ise son test sonucu olarak verilmiştir. Sonuçlardaki standart sapma değeri de, tasarlanan yöntemlerde, diğer yöntemlere nispeten daha düşüktür. Bunun nedeni, geliştirilen yöntemlerde bütün bağlantı türleri için hem benzerlik skorları hem de bağlantı sıklık değişim bilgisinin zaman içinde değişimlerinin birlikte kullanılması ve veri kümesindeki değişimden daha az etkilenilmesidir. Elde edilen sonuçlara göre bütün bağlantı türleri için düğüm benzerlik ölçütlerinin ve bağlantı sıklık değişim bilgilerinin zaman içinde değişimlerinin eşanlı analiz edilmesi bağlantı tahmini performansında artış sağlamaktadır.

5.3.3.2 Farklı bağlantı türleri için NARX modelinin gecikme değerlerinin analizi



Şekil 5.20 : Ortak-yayın bağlantı türü için gizli katmandaki nöron, gecikme ve MSE değerleri.



Şekil 5.21 : Ortak-kaynak bağlantı türü için gizli katmandaki nöron, gecikme ve MSE değerleri.

Bu bölümde ortak-yayın bağlantı türü (DBLP veri kümesi) ve ortak-kaynak (Delicious veri kümesi) bağlantı türlerinde önerilen NARX-Sim-TE-Het modeli için model parametrelerinin analizi verilmiştir. HN, ilgili bağlantı türü için modelin gizli katmandaki nöron sayısını ve D, hedef değişken olan ilgili bağlantı türü için benzerlik ölçütünün gecikme değerini göstermek üzere, Şekil 5.20 ve Şekil 5.21’de farklı HN-D konfigürasyonlarına göre MSE değerinin nasıl değiştiği verilmiştir. Şekiller en yüksek benzerlik skorunun elde edildiği 15 düğüm çiftine ait sonuçlardan oluşmaktadır. Örneğin HN9D4 gizli katmanda 9 nöronun olduğu ve hedef değişken için gecikme değerinin 4 olduğu modeli ($d_y = 4$) ifade etmektedir. Gizli katmandaki nöron sayısı 8 ve 15 arasında, gecikme değeri ise 1 ve 5 arasında en düşük MSE değerini verecek şekilde deneme yanılma yöntemiyle, optimize edilmeye çalışılmıştır. En düşük MSE değeri ortak-yayın bağlantı türü için HN9D4 konfigürasyonunda, ortak-kaynak bağlantı türü için ise HN12D2 konfigürasyonunda elde edilmiştir. Kullanılan DBLP ve Delicious gibi 2 farklı veri kümesi için gecikme değerleri incelendiğinde DBLP veri kümesine ait ortak-yayın bağlantı türü için 3-4 dönemlik gecikme değeri yeterli olurken, Delicious veri kümesine ait ortak-kaynak bağlantı türü için 2-3 dönemlik gecikme değeri yeterli olmuştur. Elde edilen zaman etkisi sonuçlarına göre Delicious veri kümesinde yakın geçmiş etkisinin ağırlığının daha fazla olduğu gözlemlenmiştir.

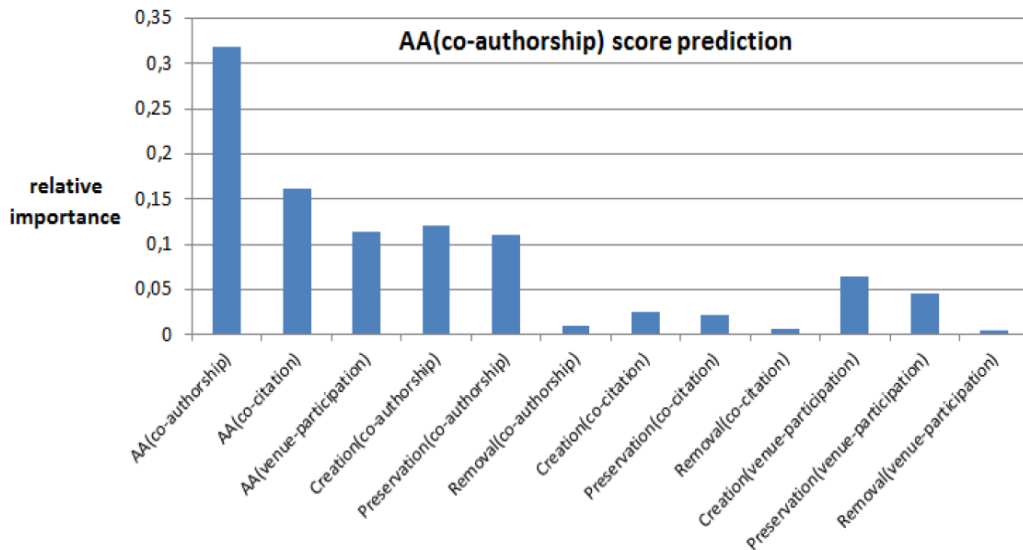
5.3.3.3 Girişler arasındaki göreceli önem dereceleri analizi

Bu bölümde, NARX-Sim-TE-Het modelindeki giriş katmanı-gizli katman ve gizli katman-çıkış katmanı arasındaki nöron bağlantılarının ağırlıkları kullanılarak girişler arasındaki göreceli önem dereceleri belirlenmiştir. Göreceli önem derecelerine göre tahmin edilmesi istenen hedef değişken ve dışsal değişkenlerin model çıktısı üzerindeki etkileri belirlenmiş ve Şekil 5.22, Şekil 5.23, Şekil 5.24, Şekil 5.25, Şekil 5.26, Şekil 5.27’de DBLP veri kümesindeki bağlantı türleri için sonuçlar verilmiştir. Sonuçlara göre farklı bağlantı türlerinin sonuç üzerindeki etkileri farklılık göstermektedir. Sonuçları sadeleştirmek için DBLP veri kümesindeki 3 farklı bağlantı türü için hesaplanan AA ve PropFlow benzerlik ölçütleri kullanılmıştır. Girişlerin göreceli önem derecelerinin belirlenmesinde bağlantı ağırlığı algoritması kullanılmıştır (İbrahim, 2013). Algoritma belirli bir giriş nöronu ile çıkış nöronu arasındaki, gizli katman nöronları aracılığıyla oluşan bağlantıların ağırlıklarını kullanarak ilgili giriş nöronunun çıkış üzerindeki etkisini göreceli olarak

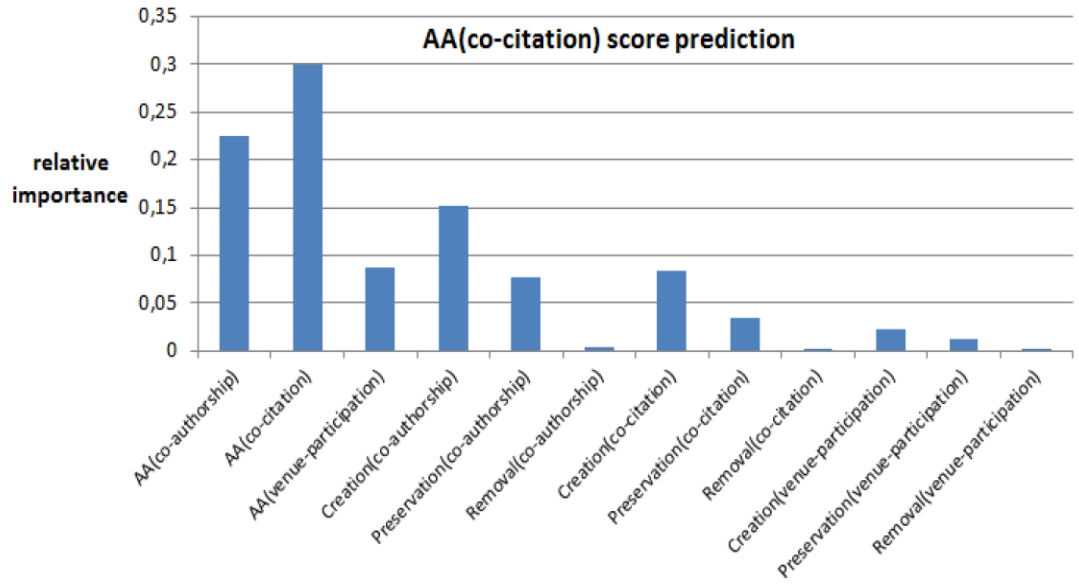
belirlemektedir. Şekil 5.22, Şekil 5.23, Şekil 5.24 ortak-yayın, ortak-atıf, ortak-konferans bağlantı türleri için AA benzerlik ölçütü kullanılarak yüzdesel olarak hesaplanan sonuçlar verilmiştir. Sonuçlara göre ilgili bağlantı türü için hesaplanan benzerlik skoru (hedef değişken) çıkış üzerinde en yüksek etkiye sahiptir. Dışsal değişkenler arasında en az etkiye kopan bağlantı sayıları sahiptir. Şekil 5.25, Şekil 5.26, Şekil 5.27’de ise benzer sonuçlar PropFlow benzerlik ölçütü kullanılarak verilmiştir.

Lokal benzerlik ölçütü AA ve global benzerlik ölçütü PropFlow için elde edilen sonuçlar karşılaştırıldığında ise farklı türdeki benzerlik ölçütlerinin çıktı üzerinde farklı etkileri olduğu görülmüştür.

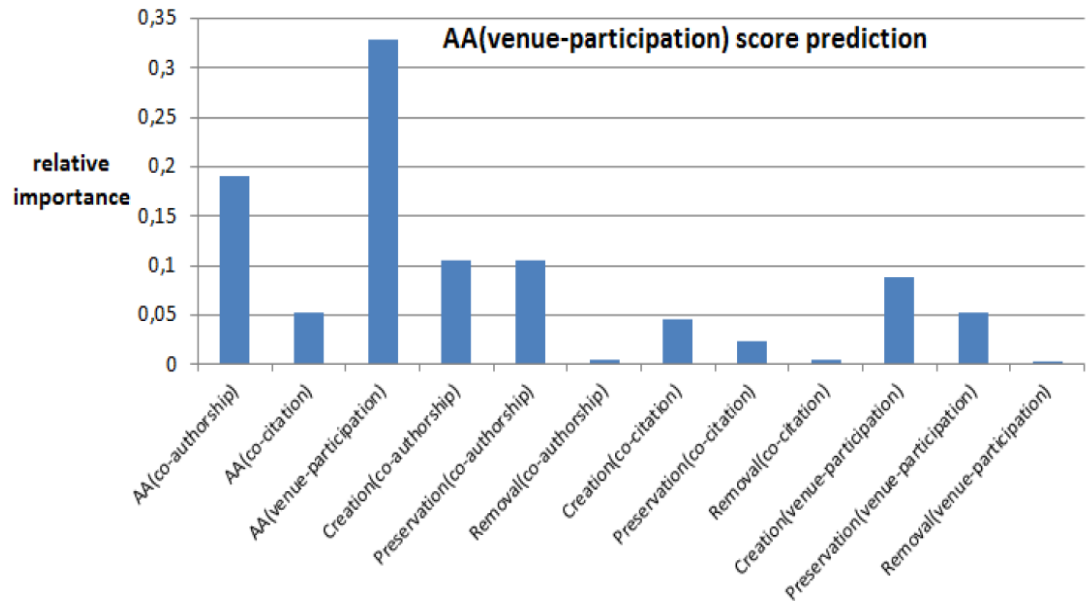
Örneğin Şekil 5.22’ye göre ortak-yayın bağlantı türüne ait AA skoru en fazla dışsal değişkenlerden ortak-atıf bağlantı türüne ait AA skorundan etkilenmiştir, Şekil 5.25’de ise ortak-yayın bağlantı türüne ait PropFlow skoru en fazla dışsal değişkenlerden ortak-konferans bağlantı türüne ait PropFlow skorundan etkilenmiştir. Bütün bağlantı türleri için dışsal değişken kopan bağlantı sayıları bilgisinin çıktı üzerinde az etkili olduğu görülmüştür. Buna göre NARX-Sim-TE-Het modelinin karmaşıklığını azaltmak ve bağlantı tahmini başarımını arttırmak için (hata oranını arttırmadan) modelin girdilerinden kopan bağlantı sayısı bilgileri çıkarılmıştır.



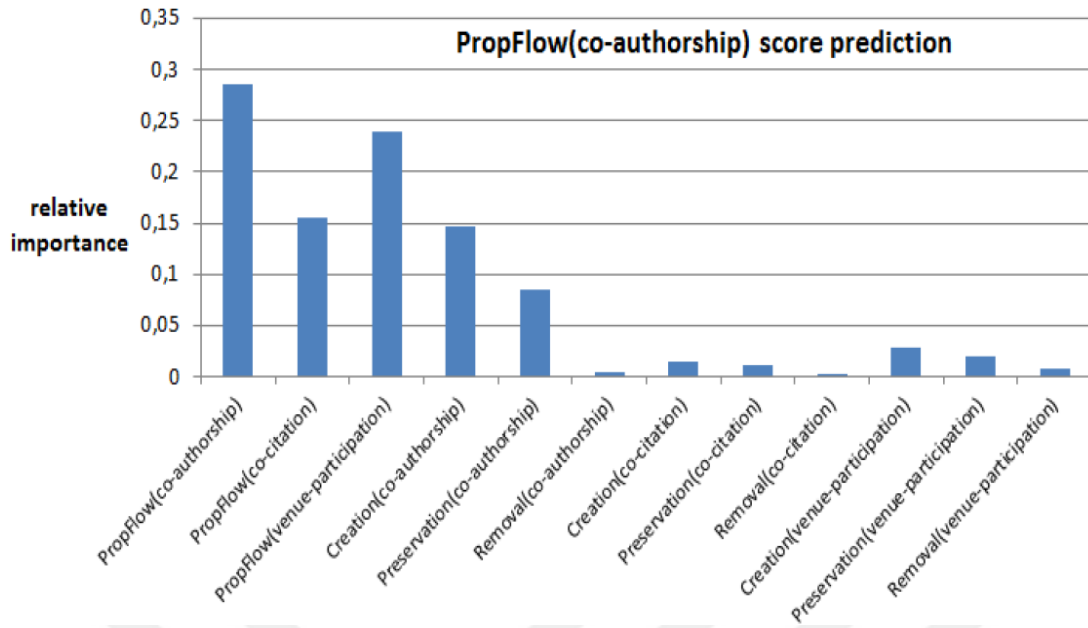
Şekil 5.22 : Ortak-kaynak bağlantı türü ve AA benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.



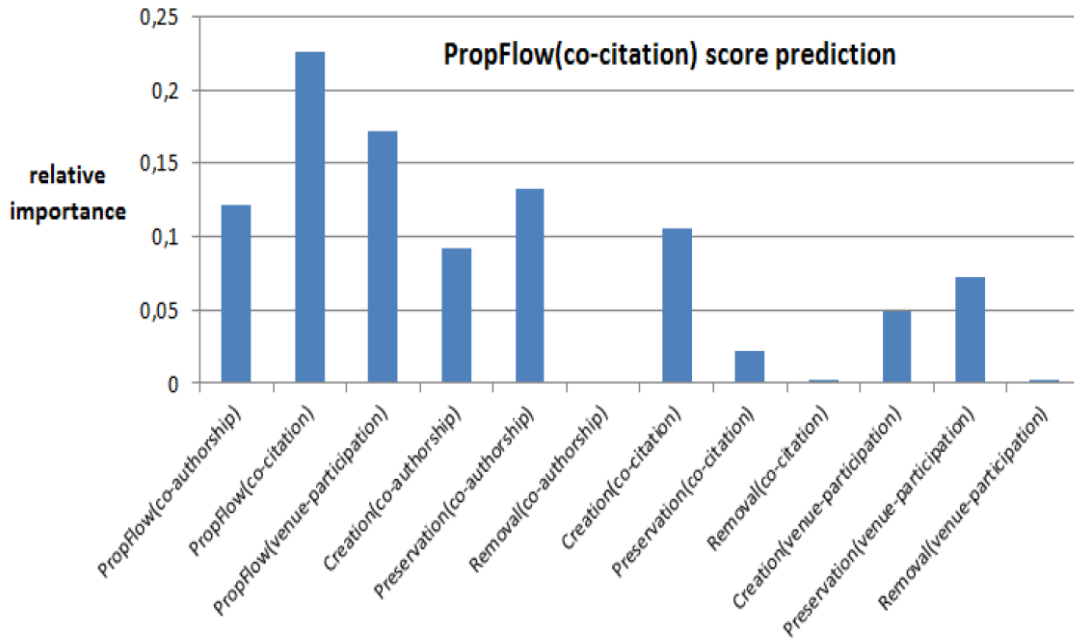
Şekil 5.23 : Ortak-atıf bağlantı türü ve AA benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.



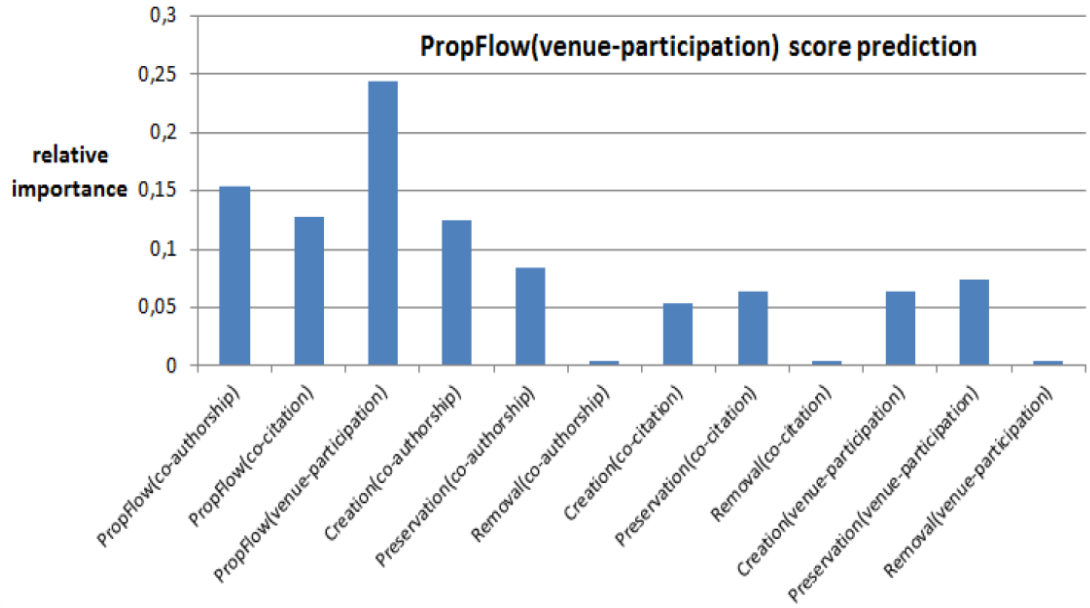
Şekil 5.24 : Ortak-konferans bağlantı türü ve AA benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.



Şekil 5.25 : Ortak-yayın bağlantı türü ve PropFlow benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.



Şekil 5.26 : Ortak-atıf bağlantı türü ve PropFlow benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.



Şekil 5.27 : Ortak-konferans bağlantı türü ve PropFlow benzerlik ölçütü için NARX-Sim-TE-Het modelinde kullanılan girişler arasındaki göreceli önem dereceleri.

5.3.3.4 Karmaşıklık analizi

Geliştirilen model, VAR modeline benzer şekilde bütün düğümler için komşuluk listesini tutan bir komşu düğümler listesi ve bütün düğüm çiftleri için ortak komşular listesi yaratarak hesaplama işlemine başlamaktadır. Tüm düğüm çiftleri için lokal benzerlik skorlarının hesaplanması, $O(N^3)$ işlemle gerçekleştirilmekte (N , düğüm sayısıdır), global benzerlik skorlarının hesaplanması ise $O(|V|^2(|V| + |E|))$ işlem ile gerçekleştirilmektedir. NARX modelinde geçmiş veriler mevcut olduğu için Seri-Paralel mimari kullanılmıştır. NARX modelinin karmaşıklığı modelin yapısına ve sinir ağındaki katman ve katmanlardaki nöron sayılarına bağlıdır. NARX modelinin karmaşıklığı $O(HN^2D)$ şeklinde verilebilir. Tahmin etmek için kullanılan NARX modelinde çıkış katmanında 1 nöron ve gizli katmanda 8-15 arası değişen nöronlar bulunduğu için ayrıca hız ve kararlılık sağlayan Levenberg-Marquardt (LMA) algoritması kullanıldığı için genel olarak karmaşıklığı benzerlik skorlarının hesaplanması oluşturmaktadır.



6. SONUÇ VE ÖNERİLER

Tez çalışmasının ilk aşamasında, zamanla değişim bilgisi kullanılarak evrimsel homojen ağlar üzerinde düğümler arasındaki bağlantıların tahmin edilmesi için çok değişkenli zaman serisi modelleri VAR-Sim ve VAR-AllSim önerilmiştir. Model ile evrimsel homojen ağlarda geçmişte ortaya çıkan düğümler arasındaki benzerlik skorları ve bu düğümler arasında meydana gelen bağlantılar eş zamanlı modellenerek, gelecekteki benzerlik skor değerleri tahmin edilmiş ve tahmin edilen bu değerlere göre hangi düğümler arasında bağlantı meydana geleceğinin tahmin edilmesi sağlanmıştır.

Öncelikle evrimsel homojen ağın her zaman dilimindeki ağ görüntüsü için düğümler arasında çeşitli ağırlıklı ve ağırlıksız benzerlik skorları hesaplanmıştır. Farklı zaman dilimlerinde hesaplanan benzerlik skorları ve bağlantı sıklık bilgisi çok değişkenli zaman serisi modeli olan VAR modeline girdi olarak verilerek gelecekteki benzerlik skorları tahmin edilmiştir. Önerilen modelin başarımı literatürdeki çeşitli tek değişkenli zaman serisi modelleri ARIMA, MA, RW ve zaman bilgisi kullanmayan statik yöntemler ile karşılaştırılmış ve %5-15 arası iyileşme sağladığı görülmüştür. Ayrıca önerilen model ile mevcut statik yöntemlerin zaman bilgisini kullanamama ve zaman serileri yöntemlerinin daha önce örneği görülmemiş bağlantıların tahmin edilememesi sorunlarına çözüm getirilmiştir.

Tez çalışmasının ikinci aşamasında, VAR-Sim ve VAR-AllSim modelleri evrimsel heterojen ağlar üzerinde düğümler arasındaki farklı tür bağlantıların tahmin edilmesi için kullanılmıştır. Benzer şekilde heterojen ağın farklı zamanlardaki görüntülerinde hesaplanan çeşitli ağırlıklı ve ağırlıksız benzerlik skorları ve bağlantı sıklık bilgisi eşzamanlı modellenmiş ve bütün bağlantı türlerine ait gelecekteki benzerlik skorları tahmin edilmiştir. Gerçeklenen model ile zaman bilgisinin etkin şekilde kullanılması, tahmin edilecek bağlantının gözlenmiş olması kabulünün ortadan kaldırılması, düğüm çiftleri arasındaki bağlantıların ve benzerlik ölçütlerinin korelasyonunun kullanılması gerçekleşmiştir. VAR modeli ile farklı bağlantı türleri için oluşturulan birden fazla zaman serileri arasındaki ilişkiler yorumlanarak mevcut heterojen bağlantı tahmini yöntemlerinin eksikliklerinin giderilmesi sağlanmıştır. Ayrıca tezin

bu aşamasında, önerilen VAR-AllSim modeli ile en iyi sonucun elde edildiği lokal ölçütler ve global ölçütler bütün bağlantı türleri için eşzamanlı modellenmiştir. Önerilen modellerin başarımı her bağlantı türününün ayrı olarak ele alındığı tek değişkenli zaman serisi modelleri ile (ARIMA-Sim ve ARIMA-Occ) karşılaştırılmış ve performansta %7-11 arası gelişme sağladığı görülmüştür.

Tez çalışmasının son aşamasında, yine evrimsel heterojen ağlar üzerinde dinamik ve döngülü sinir ağı modelleri (NARX-Sim, NARX-Sim-TE, NARX-Sim-TE-Het) önerilmiştir. VAR modelinden farklı olarak NARX modeline girdi olarak farklı bağlantı türleri için hesaplanan benzerlik skorları ve bağlantı sıklık bilgisindeki değişimler (oluşan, korunan, kopan bağlantı sayıları) verilerek eş zamanlı olarak ele alınmıştır.

Önerilen NARX modeli ile, zaman serisi modellerinde olduğu gibi serinin durağan olmasını önşart olarak kabul edilmemesi ve seriyi oluşturan veriler arasında doğrusal bir ilişkinin olduğunun varsayılmaması ile benzerlik ölçütlerinin tahmininde zaman serilerine göre farklı bir katkı sağlanmıştır. Gerçeklenen modelde ayrıca NARX sinir ağındaki nöronlar arasındaki bağlantı ağırlıkları kullanılarak girişler arasındaki göreceli önem dereceleri heterojen ağdaki farklı bağlantı türleri için belirlenmiştir.

Önerilen modeller iki farklı veri kümesi üzerinde test edilmiştir. Dengesiz veri kümelerinde de iyi sonuç vermesi nedeniyle başarımı ölçmek için AUC kullanılmıştır. Önerilen modellerin büyük veri kümelerinde ve çevrimiçi sistemlerde kullanılması gelecek çalışmalar olarak planlanmıştır. Ayrıca gelecek çalışma olarak sinir ağı modellerinden derin öğrenme yöntemlerinin kullanılması planlanmıştır.

Bağlantı tahmini modellerinin karmaşıklığı incelendiğinde, modellerin karmaşıklığını benzerlik skorlarının hesaplanması oluşturmaktadır. Benzerlik skorlarının hesaplama maliyetinin azaltılması için planlanan bir diğer gelecek çalışma da sosyal ağların çeşitli topolojik özelliklerine göre hangi benzerlik ölçütünün kullanılacağına karar verilmesidir. Örneğin, sosyal ağların demetleme katsayısı, yoğunluğu, çapı vb. topolojik özellikleri ile lokal ve global benzerlik ölçütlerinin başarımları arasında korelasyon olup olmadığı incelenerek hangi özelliğe sahip ağlarda hangi ölçütler kullanılmalı problemine çözüm getirilmesi gelecek çalışma olarak planlanmıştır.

KAYNAKLAR

- Abbasi, A.** (2016). A longitudinal analysis of link formation on collaboration networks. *Journal of Informetrics*, 10(3), 685-692
- Abbasi, A., Altmann, J., ve Hossain, L.** (2011). Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5(4), 594-607.
- Abbasi, A., Hossain, L., ve Leydesdorff, L.** (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, 6(1), 403-412.
- Abbasi, A., Wigand, R., ve Hossain, L.** (2014). Measuring social capital and its influence on individual performance. *Library & Information Science Research*, 36(1), 66-73.
- Adamic, L. A. ve Adar, E.** (2003). Friends and neighbors on the web. *Social Networks*, 25(3), 211-230.
- Ahuja, G., Soda, G., ve Zaheer, A.** (2012). The genesis and dynamics of organizational networks. *Organization Science*, 23(2), 434-448.
- Akaike, H.** (1973). Maximum Likelihood Estimation of Gaussian Autoregressive Moving Average Models. *Biometrika*, 60, 255-265.
- Akdi, Y.** (2012). *Zaman serileri analizi (Birim kökler ve kointegrasyon)*. Gazi Kitapevi, Ankara
- Akgül, I.** (2003). *Zaman Serilerinin Analizi ve Arima Modelleri*. Der Yayınları, İstanbul
- Albert, R. ve Barabási, A. L.** (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74, 47-97.
- Ardalani, M. ve Zolfaghari, S.** (2010). Chaotic time series prediction with residual analysis method using hybrid Elman-NARX neural networks. *Neurocomputing*, 73(13), 2540-2553.
- Barabasi, A. L.** (2016). *Network Science*. Cambridge University Press, United Kingdom
- Barabasi, A. L. ve Albert, R.** (1999). Emergence of scaling in random networks. *Science*, 286, 509-512.
- Barabási, A.L. ve Bonabeau, E.** (2003). Scale-Free Networks. *Scientific American*, 288(5), 50-9.
- Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A. ve Vicsek, T.** (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311, 590-614.

- Box, G. ve Jenkins, G.M.** (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.
- Bozkurt, H. Y.** (2013). *Zaman Serileri Analizi*. Bursa, Türkiye: Ekin Yayınevi.
- Brin, S. ve Page, L.** (1998). The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the seventh international conference on World Wide Web, Australia*, 107-117.
- Burt, R. S.** (1992). *Structural holes: the social structure of competition*. Cambridge, MA: Harvard University Press.
- Cannistraci, C. V., Alanis-Lobato, G., ve Ravasi, T.** (2013). *Scientific reports* 3, 1613.
- Cao, B., Kong, X. ve Yu, P. S.** (2014). Collective Prediction of Multiple Types of Links in Heterogeneous Information Networks. *IEEE International Conference on Data Mining, Shenzhen*, 50-59.
- Chatterjee, S., Nigam, S. ve Singh, J.B.** (2012). Software fault prediction using Nonlinear Autoregressive with eXogenous Inputs (NARX) network. *Appl Intell.*, 37, 121.
- Davis, D., Lichtenwalter, R. ve Chawla, N. V.** (2011). Multi-relational Link Prediction in Heterogeneous Information Networks. *ASONAM '11 Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining, IEEE Computer Society, Washington, DC*.
- Davis, D., Lichtenwalter, R. ve Chawla, N. V.** (2013). Supervised methods for multi-relational link prediction. *Social Network Analysis and Mining*, 3(2), 127-141.
- Deljac, Ž., Randić, M. ve Krčelić, G.** (2016). A Multivariate Approach to Predicting Quantity of Failures in Broadband Networks Based on a Recurrent Neural Network. *J. Netw. Syst. Manage.*, 24(1), 189-221.
- DiPrete, T. A., ve Eirich, G. M.** (2006). Cumulative advantage as a mechanism for inequality: a review of theoretical and empirical developments. *Annual Review of Sociology*, 32, 271-297.
- Engle, R. ve Granger, C.** (1987). Co-integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 55(2), 251-76.
- Erdős, P. ve Rényi, A.** (1959). On Random Graphs. *Publications Mathematicae*, 6, 290.
- Freeman, L. C.** (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1(3), 215-239.
- Granovetter, M.** (1983). The strength of weak ties: A network theory revisited. *Sociological Theory* 1, 201-233.
- Gulati, R.** (1995). Social structure and alliance formation patterns: a longitudinal analysis. *Administrative Science Quarterly*, 40, 619-652.

- Gunes, I., Gunduz-Oguducu, S. ve Cataltepe, Z.** (2015). Link prediction using time series of neighborhood-based node similarity scores. *Data Mining and Knowledge Discovery*, 1-34.
- Gürsakal, N.** (2009). *Sosyal Ağ Analizi*. Bursa, Türkiye: Dora Yayınları.
- He, J., Bailey, J. ve Zhang, R.** (2014). Exploiting Transitive Similarity and Temporal Dynamics for Similarity Search in Heterogeneous Information Networks. *Database Systems for Advanced Applications, Lecture Notes in Computer Science* Volume 8422, 141-155
- Hsu, W. H., King, A., Paradesi, M.S.R., Pydimarri, T. ve Weninger, T.** (2006). Collaborative And Structural Recommendation Of Friends Using Weblog-Based Social Network Analysis. *Proc. of Computational Approaches to Analyzing Weblogs - AAAI 2006 Technical Report SS-06-03*, pp. 55-60.
- Huang, Z. ve Lin, D. K. J.** (2009). The time-series link prediction problem with applications in communication surveillance. *INFORMS J. Comput.* 21, 286.
- Ibrahim, O.M.** (2013). A comparison of methods for assessing the relative importance of input variables in artificial neural networks. *Journal of Applied Sciences Research*, Vol. 9 Issue 11, p5692
- Johnson, R.A., Yang, Y., Aguiar, E., Rider, A. ve Chawla, N.V.** (2013). ALIVE: a multi-relational link prediction environment for the healthcare domain. *Proceeding PAKDD'12 Proceedings of the 2012 Pacific-Asia conference on Emerging Trends in Knowledge Discovery and Data Mining Pages*, 36-46 Springer-Verlag Berlin, Heidelberg.
- Junior, J.M.P.M. ve Barreto, G. A.** (2008). Long-term time series prediction with the NARX network: An empirical evaluation. *Neurocomputing*, 71(16), 3335-3343.
- Leskovec, J., Kleinberg, J. ve Faloutsos, C.** (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. *In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD '05)*. 177-187, ACM, New York, NY, USA.
- Katz, L.** (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39–43.
- Kilian, L. ve Chang, P.L.** (2000). How Accurate are Confidence Intervals for Impulse Responses in Large VAR Models. *Economics Letters*, Vol. 69, 299- 307.
- Lee, J.B. ve Adorna, H.** (2012). Link Prediction in a Modified Heterogeneous Bibliographic Network. *ASONAM, IEEE Computer Society*, page 442-449.
- Levin, D.Z. ve Cross, R.** (2004). The Strength of Weak Ties You Can Trust: The Mediating Role of Trust in Effective Knowledge Transfer. *Manage. Sci.*, 50(11), 1477-1490.
- Liben-Nowell, D. ve Kleinberg, J. M.** (2003). The link prediction problem for social networks. *CIKM*, 556-559.

- Lichtenwalter, R. N., Lussier, J. T. ve Chawla, N. V.** (2010). New perspectives and methods in link prediction. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, KDD '10, pp 243-252
- Liu F., Liu B., Sun C., Liu M. ve Wang X.** (2013) Deep Learning Approaches for Link Prediction in Social Network Services. In: *Lee M., Hirose A., Hou ZG., Kil R.M. (eds) Neural Information Processing. ICONIP*. Lecture Notes in Computer Science, vol. 8227. Springer, Berlin, Heidelberg
- Liu, W. ve Lü, L.** (2010). Link prediction based on local random walk. *Europhys. Lett.*, vol. 89, 58007.
- Lutkepohl, H.** (1982). Non-Causality Due to Omitted Variables. *Journal of Econometrics*, 19, 367-378.
- Lü, L. ve Zhou, T.** (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, vol. 390(6), pp. 1150-1170.
- Merton, R. K.** (1968). The Matthew effect in science. *Science*, 159(3810), 56–63.
- Stanley, M.** (1967). The Small World Problem. *Psychology Today*, 1(1).
- Munasinghe, L. ve Ichise, R.** (2012). Time Score: A New Feature for Link Prediction in Social Networks. *IEICE Transactions on Information and Systems* 95(3), 821-828.
- Murata, T. ve Moriyasu, S.** (2007). Link Prediction of Social Networks Based on Weighted Proximity Measures. *Web Intelligence*, 85-88.
- Newman, M.** (2001a). Clustering and preferential attachment in growing networks. *Physical Review E*, 64, 025102.
- Newman, M.** (2001b). Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E*, 64, 016131.
- Newman, M.** (2001c). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64, 016132.
- Newman, M.** (2003a). The structure and function of complex networks. *SIAM Review*, 45, 167-256.
- Ou, Q., Jin, Y.D., Zhou, T. ve Wang, B.H. and Yin, B.Q.** (2007). Powerlaw strength-degree correlation from resource-allocation dynamics on weighted networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 75(2), 021102.
- Özdemir, A. F., Yıldıztepe, E. ve Binar, M.** (2010). İstatistiksel Yazılım Geliştirme Ortamı: R. 12. *Akademik Bilişim Konferans Bildirileri*, Muğla Üniversitesi.
- Pindyck, R.S. ve Rubinfeld, D.L.** (1991). *Econometric Models and Economic Forecasts*. MacGraw-Hill, New York, First Edition.
- Salton, G. ve McGill, M. J.** (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

- Soares, P. R.S. ve Prudêncio, R.B.C.** (2012). Time Series Based Link Prediction. *IJCNN the 2012 International Joint conferences*, pp. 1-7, IEEE.
- Soares, P. R.S. ve Prudêncio, R.B.C.** (2013). Proximity measures for link prediction based on temporal events. *Expert Systems with Applications*, vol. 40, pp. 6652-6660.
- Sun, Y., Barber, R., Gupta, M., Aggarwal, C.C. ve Han, J.** (2011). Co-author relationship prediction in heterogeneous bibliographic networks. *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011*, IEEE Computer Society, Washington, DC.
- Tan, P.-N., Steinbach, M. ve Kumar, V.** (2005). *Introduction to Data Mining*. ISBN 0-321-32136-7.
- Tong, H., Faloutsos, C. ve Pan, J.** (2006). Fast Random Walk with Restart and Its Applications. *Sixth International Conference on Data Mining (ICDM'06)*, Hong Kong, pp. 613-622.
- Url-1** <<http://sourceforge.net/projects/ujmp/>>, alındığı tarih: 12.10.2014.
- Url-2** <<http://stax.codehaus.org/>>, alındığı tarih: 13.10.2014.
- Url-3** <<https://www.r-project.org/>>, alındığı tarih: 12.10.2014.
- Url-4** <http://konect.uni-koblenz.de/dblp_coauthor>, alındığı tarih: 13.10.2014.
- Url-5** <<https://del.icio.us/>>, alındığı tarih: 13.10.2014.
- Watts, D. J., ve Strogatz, S. H.** (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684), 440–442.
- Wu, Z., Lin, Y., Wang, J. ve Gregory, S.** (2016). Link prediction with node clustering coefficient. *Physica A: Statistical Mechanics and its Applications*, 452, 1-8.
- Yang, Y., Chawla, N.V., Sun, Y. ve Han, J.** (2012). Link prediction in heterogeneous networks: Influence and time matters.
- Yao, L., Wang, L., Pan, L. ve Yao, K.** (2016). Link Prediction Based on Common-Neighbors for Dynamic Social Network, *Procedia Computer Science*, Volume 83, 2016, Pages 82-89.
- Zhou, T., Lü, L. ve Zhang, Y. C.** (2009). Predicting missing links via local information. *Eur. Phys. J. B.*, 71, 623.
- Zhou, T., Ren, J., Medo, M. ve Zhang Y. C.** (2007). Bipartite network projection and personal recommendation. *Phys. Rev. E.* 76, 046115.



EKLER

EK A: Terimlerin İngilizce Karşılıkları



EK A

Çizelge A.1 : Terimler ve ingilizce karşılıkları.

Terim	Karşılığı (İngilizce)
Sosyal Ağlar	Social Networks
Düğüm	Node
Kenar	Edge
Derece	Degree
Küçük Dünya	Small World
Vektör Otoregresyon	Vector Autoregression

ÖZGEÇMİŞ

Ad-Soyad : Alper ÖZCAN
Doğum Tarihi ve Yeri : 13/09/1984, Antalya
E-posta : alper.ozcan@gmail.com

ÖĞRENİM DURUMU:

- **Lisans** : 2007, İTÜ, Elektrik-Elektronik Fak., Bilgisayar Mühendisliği
- **Yüksek Lisans** : 2010, İTÜ, Bilgisayar Mühendisliği, Bilgisayar Mühendisliği

MESLEKİ DENEYİM VE ÖDÜLLER:

- 2008-2009 yılları arasında TÜBİTAK MAM Araştırmacı olarak çalıştı
- 2009-2010 yılları arasında Turkcell Teknoloji Yazılım mühendisi olarak çalıştı.
- 2011-Halen Softtech Şirketinde Ekip Lideri olarak çalışmaktadır.

DOKTORA TEZİNDEN TÜRETİLEN YAYINLAR, SUNUMLAR VE PATENTLER:

- **Özcan, A.** ve Gündüz-Öğüdücü, Ş. (2017). Link Prediction in Evolving Heterogeneous Networks using the NARX Neural Networks. *Knowledge and Information Systems*.
- **Özcan, A.** ve Gündüz-Öğüdücü, Ş. (2017). Temporal Link Prediction Using Time Series of Quasi-Local Node Similarity Measures. *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, CA, pp. 381-386.
- **Özcan, A.** ve Gündüz-Öğüdücü, Ş. (2015). Multivariate Temporal Link Prediction in Evolving Social Networks, *International Journal of Computer & Information Science (IJCIS)*, vol. 16, 3, 2015.
- **Özcan, A.** ve Gündüz-Öğüdücü, Ş. (2015). Multivariate temporal Link Prediction in evolving social networks. *IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS)*, Las Vegas, NV, pp. 185-190.

