

**T.C.**  
**GEBZE YÜKSEK TEKNOLOJİ ENSTİTÜSÜ**  
**MÜHENDİSLİK ve FEN BİLİMLERİ ENSTİTÜSÜ**

**ŞÜPHELİ AĞLARDA TOPLULUK TANIMA**

**ARZU KAKIŞIM**  
**YÜKSEK LİSANS TEZİ**  
**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**GEBZE**  
**2013**

**T.C.**  
**GEBZE YÜKSEK TEKNOLOJİ ENSTİTÜSÜ**  
**MÜHENDİSLİK ve FEN BİLİMLERİ ENSTİTÜSÜ**

**ŞÜPHELİ AĞLARDA TOPLULUK TANIMA**

**ARZU KAKIŞIM**  
**YÜKSEK LİSANS TEZİ**  
**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**DANIŞMANI**  
**PROF. DR. İBRAHİM SOĞUKPINAR**

**GEBZE**  
**2013**



**GEBZE YÜKSEK  
TEKNOLOJİ ENSTİTÜSÜ**

## **YÜKSEK LİSANS JÜRİ ONAY FORMU**

GYTE Mühendislik ve Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 17/06/2013 tarih ve 2013/32 sayılı kararıyla oluşturulan jüri tarafından 26/08/2013 tarihinde tez savunma sınavı yapılan Arzu KAKIŞIM'ın tez çalışması Bilgisayar Mühendisliği Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

### **JÜRİ**

ÜYE

(TEZ DANIŞMANI) : Prof. Dr. İbrahim SOĞUKPINAR

ÜYE

: Doç. Dr. Mehmet GÖKTÜRK

ÜYE

: Yrd. Doç. Dr. Serdar Süer ERDEM

### **ONAY**

GYTE Mühendislik ve Fen Bilimleri Enstitüsü Yönetim Kurulu'nun ..... tarih ve ...../..... sayılı kararı.

**İMZA/MÜHÜR**

## ÖZET

Son yıllarda sosyal ağların kullanımının hızlı bir şekilde artması ile sosyal ağlarda topluluk tanıma konusu önem kazanmaya başlamıştır. Sosyal dokümanların ve bağlantıların analiz edilerek gizil toplulukların keşfedilmesine yönelik birçok yöntem geliştirilmiştir. Bu çalışmada, suç ağı olduğu kanıtlanmamış olası şüpheli ağların analiz edilmesi için bir üretici olasılıklı topluluk tanıma modeli geliştirilmiştir. Bu model, klasik yöntemlerdeki bağlantı bilgisinin keşfinin yanında sosyal ağdaki aktörlerin ilgilendiği konulara dair içerik analizi yapmaktadır. Bir sosyal ağdaki kullanıcılara ait yönlü mesaj verisi kullanılarak, benzer konuları konuşan ağ üyeleri aynı topluluk altında kümelenmeye çalışılmıştır. Bu çalışmada geliştirilen model için Topluluk-Konu (TK) modeli temel alınmıştır. TK modeli, ağın bağlantı bilgisini kullanmadığı için topolojik olarak ayırık topluluklar üretmektedir. Bu eksiklik, TK modeline bağlantı bilgisinin eklenmesi ile çözülmeye çalışılmıştır. Enron şirketine ait e-posta verileri üzerinde yapılan testler ile hem içerik hem de topolojik olarak benzer kullanıcıların başarılı bir şekilde bir topluluk altında kümelendiği gözlenmiştir.

**Anahtar Kelimeler:** Sosyal Ağ Analizi, Şüpheli Ağı, Konu Modeli, Topluluk Tanıma.

## SUMMARY

In recent years, with the rise of the use of social networks, the issue of community detection in social networks has come into prominence. Many approaches have been developed to extract latent communities analyzing social documents and connections. In this study, a generative probabilistic community detection model is developed to analyze suspicious networks which are not proven as criminal networks. In addition to the discovery of connection information in classic methods, this model makes content analysis about topics interested by actors in the social network. By the use of directive message information of users in a network, network members who talk about the same topics, are tried to be grouped in the same community. Community-topic model is used as base for the model developed in this study. Community-topic model produces topologically diverse communities as it does not use connection information of the network. This deficiency has been tried to be solved with the addition of connection information to the community-topic model. With the tests made on the e-mail information belonging to Enron company, it is observed that topically and topologically similar users are successfully grouped in a single community.

**Keywords: Social Network Analysis, Suspect Network, Topic Model, Community Detection.**

## TEŐEKKÖR

Bu tez alıřmasının danıřmanlıđını yÖrÖten, benden hibir zaman yardımını ve bilgisini esirgemeyen deđerli hocam Prof. Dr. İbrahim SOĐUKPINAR’a,

alıřmam sÖresince bana destek olan arkadaşlarım Okyanus TULGAR’a ve Sevda VATANSEVER’e,

Bana her zaman destek olan ve hibir fedakârlıktan kaçınmayan kıymetli aileme ve eřim Can KAKIŐIM’a sonsuz teőekkÖrlerimi sunarım.

# İÇİNDEKİLER

|                                           | <b><u>Sayfa</u></b> |
|-------------------------------------------|---------------------|
| ÖZET                                      | iv                  |
| SUMMARY                                   | v                   |
| TEŞEKKÜR                                  | vi                  |
| İÇİNDEKİLER                               | vii                 |
| SİMGELER ve KISALTMALAR DİZİNİ            | ix                  |
| ŞEKİLLER DİZİNİ                           | x                   |
| TABLOLAR DİZİNİ                           | xi                  |
| <br>                                      |                     |
| 1. GİRİŞ                                  | 1                   |
| 2. TEMEL BİLGİLER VE İLGİLİ ÇALIŞMALAR    | 6                   |
| 2.1. Çizge Kuramı                         | 6                   |
| 2.1.1. Çizgenin Tarihi                    | 7                   |
| 2.1.2. Çizge Özellikleri                  | 8                   |
| 2.2. Suçlu ve Şüpheli Ağları              | 9                   |
| 2.2.1. Suçlu ve Şüpheli Ağları Analizi    | 11                  |
| 2.3. Bayes Olasılığı ve Konu Modelleri    | 15                  |
| 2.3.1. Bayes Olasılığı                    | 15                  |
| 2.3.2. Frekans Olasılığı                  | 16                  |
| 2.3.3. Üretken Modeller                   | 17                  |
| 2.3.4. Markov Zinciri Monte Carlo Yöntemi | 19                  |
| 2.3.5. Gizil Dirichlet Tahsisi            | 21                  |
| 2.3.6. Yazar-Konu Modeli                  | 24                  |
| 2.3.7. Gibbs Örnekleme                    | 26                  |
| 2.3.8. Yazar-Alıcı-Konu Modeli            | 29                  |
| 2.4. Topluluk Tanıma Modelleri            | 30                  |
| 2.4.1. Topluluk-Kullanıcı-Konu Modeli     | 31                  |
| 2.4.2. Topluluk-Yazar-Alıcı-Konu Modeli   | 32                  |
| 2.4.3. Topluluk-Konu Modeli               | 34                  |
| 3. GELİŞTİRİLEN MODEL                     | 37                  |
| 3.1. Yazar-Alıcı-Topluluk-Konu Modeli     | 37                  |

|                                              | <b><u>Sayfa</u></b> |
|----------------------------------------------|---------------------|
| 3.2. Analiz Yöntemleri                       | 39                  |
| 3.2.1. Kullback Leibler Yöntemi              | 40                  |
| 3.2.2. Normalleştirilmiş Ortak Bilgi Yöntemi | 40                  |
| 3.2.3. Modülerlik Yöntemi                    | 41                  |
| 4. DENEYSEL ÇALIŞMALAR                       | 42                  |
| 4.1. Veri Kümesi                             | 42                  |
| 4.2. Deneyler                                | 42                  |
| 4.2.1. Topluluk ve Konu Analizi              | 43                  |
| 4.2.2. Bağlantı Analizi                      | 46                  |
| 4.2.3. Çalışma Süresi Analizi                | 49                  |
| 5. SONUÇ VE ÖNERİLER                         | 52                  |
| <br>KAYNAKLAR                                | <br>53              |
| ÖZGEÇMİŞ                                     | 57                  |

## SİMGELER ve KISALTMALAR DİZİNİ

### Simgeler ve Açıklamalar

### Kisaltmalar

|              |   |                                                    |
|--------------|---|----------------------------------------------------|
| $C$          | : | Topluluk sayısı                                    |
| $D$          | : | $M$ sayıda dokümandan oluşan yapı                  |
| $M$          | : | Doküman sayısı                                     |
| $N$          | : | Bütün dokümanlar içerisindeki toplam kelime sayısı |
| $N_d$        | : | $d$ dokümanına ait kelime sayısı                   |
| $T$          | : | Konu sayısı                                        |
| $V$          | : | Sözlükteki kelime sayısı                           |
| $\mathbf{w}$ | : | $N$ sayıda kelime içeren dokümana ait vektör       |
| $w$          | : | Kelime                                             |
| GAI          | : | Gizli Anlamsal İndeksleme                          |
| GDT          | : | Gizli Dirichlet Atama                              |
| NOB          | : | Normalleştirilmiş Ortak Bilgi                      |
| OGAI         | : | Olasılıklı Gizli Anlamsal İndeksleme               |
| TF-TDF       | : | Terim Frekansı-Ters Doküman Frekansı               |
| TK           | : | Topluluk-Konu                                      |
| TKK          | : | Topluluk-Kullanıcı-Konu                            |
| TYAK         | : | Topluluk-Yazar-Alıcı-Konu                          |
| YAK          | : | Yazar-Alıcı-Konu                                   |
| YATK         | : | Yazar-Alıcı-Topluluk-Konu                          |
| YK           | : | Yazar-Konu                                         |

## ŞEKİLLER DİZİNİ

| <b><u>Sekil No:</u></b>                                                | <b><u>Sayfa</u></b> |
|------------------------------------------------------------------------|---------------------|
| 2.1: Könisberg'in yedi köprüsü.                                        | 7                   |
| 2.2: Könisberg probleminin çizge yapısı ile gösterimi.                 | 7                   |
| 2.3: Krebs'in 19 uçak korsanına ait terörist ağı.                      | 9                   |
| 2.4: a) Netmap'a ait ağ görüntüsü, b) Xanalyst'e ait ağ görüntüsü.     | 10                  |
| 2.5: Yapısal ağ analiz yaklaşımı, a) Tüm ağ görüntüsü, b) Alt gruplar. | 11                  |
| 2.6: GDT modelinin plaka gösterimi.                                    | 22                  |
| 2.7: YK modelinin plaka gösterimi.                                     | 24                  |
| 2.8: YAK modelinin plaka gösterimi.                                    | 29                  |
| 2.9: TKK modelinin plaka gösterimi.                                    | 31                  |
| 2.10: TYAK modelinin plaka gösterimi.                                  | 33                  |
| 2.11: TK modelinin plaka gösterimi.                                    | 35                  |
| 3.1: YATK modelinin plaka gösterimi.                                   | 37                  |
| 4.1: Topluluk-konu dağılımı.                                           | 44                  |
| 4.2: Kullanıcı-topluluk dağılımı.                                      | 45                  |
| 4.3: 1. yazar için alıcı-topluluk dağılımı.                            | 47                  |
| 4.4: 3. topluluğa ait ağ görüntüsü, eşikdeğeri $\geq 0,5$ .            | 48                  |
| 4.5: 3. topluluğa ait ağ görüntüsü, eşikdeğeri $\geq 0,65$ .           | 49                  |
| 4.6: 3. topluluğa ait ağ görüntüsü, eşikdeğeri $\geq 0,9$ .            | 49                  |
| 4.7: YATK ve TK modeline ait çalışma süreleri.                         | 50                  |

## TABLÖLAR DİZİNİ

| <b><u>Tablo No:</u></b>                           | <b><u>Sayfa</u></b> |
|---------------------------------------------------|---------------------|
| 4.1: Enron yapısına ait üretilen konular.         | 43                  |
| 4.2: YATK ve TK modele ait ortalama KL değerleri. | 45                  |
| 4.3: YATK ve TK model için NOB değerleri.         | 46                  |

# 1. GİRİŞ

Toplumdaki güveni ve asayişini sağlamak için bir suç işlemiş olan kişilerin başka bir suç işlemelerini engellemek adına hızlı bir şekilde yakalanmaları gerekmektedir. Emniyeti sağlamak üzere çalışan birimler teknolojinin gelişimi ile birlikte suçlulara ait bilgileri veri tabanlarında tutar ve bu bilgileri kullanarak başka bir olayın çözümlenmesini sağlarlar. Teknolojinin bu gelişimi emniyet birimlerine hız kazandırmaktadır. Bir soruşturma aşamasında, olayın nedenine dair bilgi edinmek ve bu bilgilere göre suçluları yakalamak için emniyet birimleri görgü tanığı ifadeleri ve olaya ait delilleri analiz ederler. Analiz sırasında elde edilen bütün veriler kamu güvenliği kurumlarının veri tabanlarında tutulur ve suçlulara ait profiller yaratılır. Bu profiller, gelecekte işlenecek olası suçları çözümlmek için kullanılmaktadır. Kamu güvenliğini sağlayan kurumlar için bazı yazılımlar yapılmakta ve akademik olarak yayınlar hazırlanmaktadır. Örneğin, Bayes ağ modeli kullanılarak, görgü tanıklarının ifadelerine ve delillere göre olası suçlulara ait yaş, meslek ve cinsiyet gibi bilgileri belli olasılıklarla bulup, farklı suç olaylarına göre farklı suç profilleri oluşturan bir sistem tasarlanmıştır [8]. Bu sistem ile şüphelilere ait suçlu olma olasılıkları hesaplanabilmektedir.

Suçlulara ait profillere ve bilinen yada tahmin edilen ilişkilere göre soruşturmayı yapan araştırmacılar olaylara ait suçlu ağları oluşturmaktadır. Bu ağlar çizge yapısını kullanılarak tasarlanır. Bir çizgede bulunan düğümler suçluları, bu düğümler arasında var olan bağlantılar ise suçlular arasındaki ilişkiyi temsil etmektedir. Klasik suçlu ağ analizi yöntemlerinde, ağa ait düğüm ve kenar bilgisini kullanılır. Örneğin, ağdaki en fazla bağlantı sayısına sahip bir düğüm ağın en aktif üyesi olarak değerlendirilir. Eğer bir olay hala soruşturma altında ise ve gerçek suçlulara henüz ulaşılamamış ise ağlar olasılıklara bağlı olmalıdır. Çünkü bu ağlardaki kişilerin suçlu yada masum oldukları bilinmemektedir. Olasılıklı bu ağ modeli şüpheli ağ olarak adlandırılır. Ağın oluşturulması ve en doğru şekilde analiz edilmesi olayın en kısa zamanda çözümlenmesi için çok önemlidir. Bu problemin çözümü için bir şüpheli ağına ait şüpheliler arasında öncelikli olarak hangi bireyin incelenmesi gerektiğine ve ağdaki suçlu bir birey teşhis edildiğinde ağın tekrar nasıl kurulması gerektiğine dair bir çalışma yapılmıştır [3]. Model inanç yayılımı algoritması baz alınarak geliştirmiştir. Bu yazar olasılıklara dayalı şüpheli ağ

üzerine yoğunlaşarak suçlu ağı konusu ile ilgili çalışma alanını genişletmiştir. Fakat çalıştığı şüpheli ağ tamamen elle hazırlanmış ve bu ağın otomatik olarak oluşturulması için bir yöntem olmadığı belirtmiştir.

Kamu güvenlik kurumları için yapılan bu tür çalışmalar suçlu ve şüpheli ağı analizi adıyla literatürdeki yerini almıştır. Bu başlık altında, suçlu ve şüpheli ağlarını topolojik olarak ya da içerik açısından analiz eden birçok farklı yöntem geliştirilmiş ve birçok farklı alt konu incelenmiştir. Polis departmanlarında bulunan verileri inceleyerek suçlular arasındaki ilişkileri ortaya çıkaran ve ağın topolojik analizini yapan bir uygulama geliştirilmiştir [9], [10]. Bu uygulama da örneğin, bir şehirdeki yada bir ülkedeki araba hırsızlarının hepsi bir ağda görüntülenmektedir. Bu suçlu ağı incelenerek çeşitli gruplaşmalar yada suçlular arasındaki hiyerarşiler tespit edilmeye çalışılmaktadır. Zaman içerisinde yeni suçlular iştirak edebilir ve suçlular arasındaki dengeler değişebilir. Bu değişimin oluşturulan ağ üzerinden takibi daha kolay olabilmektedir. Kişilerin sabit disklerinde bulunan dağınık verilere veri madenciliği teknikleri uygulayarak kişilere ait dokümanların içeriği hakkında bilgi veren bir sistem geliştirilmiştir [6]. Sistem, dokümana ait konu başlıklarını, dokümanlarda geçen telefon numaralarını, şehir isimlerini ve kişilere ait isim bilgilerini elde etmeye çalışır. Böylece, sabit disk verileri incelenen kişilerin suçlu yada şüpheli olması durumunda, bu kişilerle ilişkili diğer suçlular hakkında bilgi edinilmesi amaçlanmıştır. Bu yöntem şüphelilere ait belgelerin içerik analizini sağlamaktadır. Diğer bir uygulama [4] internet üzerinde aktif olan şüpheli sitelere ait içerik analizi yapmakta ve cihazçı sitelerde aktif olan ve şüpheli görünen topluluklar hakkında bilgi vermektedir. Bu çalışmada olasılıklı konu modeli olan Gizil Dirichlet Tahsisi (GDT) modeli [2] kullanılmıştır. Sitede var olan gruplara ait yazışmalar analiz edilerek ağda konuşulan konular listelenmiştir. Böylece tehlikeli olarak görülen konu başlıkları altında yazan kullanıcıların takibi sağlanmaktadır. Bu çalışma şüpheli ağlara ait içerik analizin önemi için iyi bir örnek teşkil etmektedir. Bu tez çalışmasında, şüpheli bir ağın içerik ve bağlantı analizi bir topluluk-konu modeli kullanılarak gerçekleştirilecektir.

Son zamanlarda, sosyal ağ sitelerinin (Facebook, Twitter, Flickr, vb.) popülaritesi hızlı bir şekilde artmaktadır. Kullanıcılar, sosyal ağların sağladığı birçok servisten yararlanmakta, yeni insanlarla tanışmakta, eski arkadaşları ile iletişime geçmekte, düşüncelerini ve ilgi alanlarını paylaşmaktadırlar. Ticaret, akademi, ve güvenlik gibi bir çok farklı alan sosyal ağ analiz konusunun gelişmesine katkıda

bulunmaktadır. Sosyal ağ analizinin en önemli problemi topluluk tanıma konusudur. Bir sosyal ağda belirli ölçütlere göre birbirine yakın olan insanların oluşturduğu öbeğe topluluk denir. Benzer bağlantılara, paylaşılan ortak konulara ve yakın konumlara göre insanlar topluluklara ayrılabilirler. Sosyal ağ birden fazla topluluğa ayırtırmak ağın analizini kolaylaştırmaktadır. Aynı topluluğa ait olan insanların ortak ilgi alanlarına sahip olduğu ön görüldüğünde kişilerin ve kişiler arasındaki bağlantıların analizi de kolaylaşmaktadır.

Topluluk tanıma üzerine yapılmış çalışmalar, genellikle üyeler arasındaki bağlantı bilgisini kullanan topoloji tabanlı topluluk analizi üzerinedir. Bunlardan en sık kullanılanlarında biri Çizge parçalama [11] yaklaşımıdır. Sosyal ağlardaki toplulukları tanımak için kullanılan önemli bir yöntemdir. Toplulukların arasındaki kenarların sayısını minimize eder. Çok kullanılan diğer bir yöntem ise Girvan Newman [12] yaklaşımıdır. Fakat bu yöntemler toplulukları topolojik olarak düzgün ayırtırmasına rağmen, kişilerin ilgi alanları ve konuştukları konuları göz ardı ettiği için içerik olarak ayrık alt topluluklar oluşturmaktadır. Bu problemin çözümü için konu modelleri kullanılmaya başlanmıştır.

Sosyal ağlardaki her bir üyenin paylaştığı veriler üyenin ilgi alanları hakkında bilgi vermektedir. Konu modelleri üyelere ait dokümanları belli konu başlıkları altında özetler. En iyi bilinen konu modeli Gizil Dirichlet Tahsisi modelidir. Bu olasılıklı model bir veri kümesine ait konuların çok katlı kelime dağılımlarını keşfeder. Bu model Yazar-Konu (YK) [1] ve Yazar-Alıcı-Konu (YAK) [13] modeli olarak geliştirilmiş ve hangi kullanıcıların keşfedilen konularla ne kadar ilgili olduğu bilgisi de öğrenilmeye çalışılmıştır. Bu modeller esas alınarak konu tabanlı topluluk tanıma yöntemleri geliştirilmiştir. Toplulukların keşfedilmesi için sosyal ağlardaki anlamsal içerik kullanılmaktadır. Sosyal ağdaki üyelerin iletişimindeki anlamsal benzerliğe göre Topluluk-Kullanıcı-Konu modeli (TKK) [5] geliştirilmiştir. Topluluklar ağ üyeleri ile, üyeler ise konu dağılımı ile modellenmektedir. Ağdaki anlamsal olarak benzer gizil toplulukları ve ilişkileri elde etmeye çalışır. Fakat, model sadece ağdaki içerik bilgisini kullanır ve ağın bağlantı yapısını göz ardı eder. Bu yapının sosyal ağlarda daha verimli kullanılması için bağlantı yapısını kullanan Topluluk-Yazar-Alıcı Konu modeli (TYAK) [14] geliştirilmiştir. Ağın bağlantı yapısını ilk kullanan YAK modelidir. Fakat bu modelde topluluk tanıma süreci yoktur. TYAK modeli için TKK ve YAK modeli birleştirilerek hibrid bir yapı elde

edilmiştir. TYAK modelinde, her dokümanı bir topluluğa ve o dokümana ait yazarlar ve alıcı kümesi de bu topluluğa tahsis edilmektedir.

Yukarıda bahsedilen topluluk modellerinde topluluklara tahsis edilen konular benzerlik taşıyabilmektedir. Toplulukların arasındaki benzerlik seviyesini azaltan bir yaklaşım olmadığı için özellikle TYAK modelinde olduğu gibi kişiler birden fazla topluluğa tahsis edilir ve keşfedilen topluluklar birbirine benzemeye başlar.

Konu dağılımlarının toplulukların keşfedilmesi sürecine nasıl katkıda bulunduğu konusu incelenerek, farklı bir yaklaşım olan Topluluk-Konu (TK) modeli geliştirilmiştir [15]. TK modeli bireylerin konu dağılımları arasındaki benzerliğe göre kişileri farklı topluluklara tahsis eder. Bir topluluk altında kümelenmiş kişilerin diğer yöntemler ile karşılaştırıldığında çok daha doğru bir şekilde benzer ilgi alanlarına ve ortak konulara sahip olduğu gözlenmiştir. Bu durum yüksek kalitede topluluklar elde edilmesini sağlamaktadır. Girvan Newman yönteminin aksine bu yöntem konu benzerliği yüksek fakat topolojik olarak ayırık topluluklar oluşturmaktadır.

Literatürdeki şüpheli ağı ve topluluk tanımaya ait çalışmalar incelenerek, bu tezde yapılmak istenen bir sosyal ağdaki şüpheli toplulukların keşfedilmesi problemi için en uygun yöntemin TK modeli olduğu görülmüştür. Çünkü suç ile ilişkisi olduğu düşünülen konu başlıkları altında toplanmış üyelerin en doğru tahmini konu benzerliklerine göre ayrılmış olan toplulukların üretilmesi ile elde edilebilir. Fakat bir şüpheli topluluk için şüpheliler arasındaki bağlantı analizi önemlidir. TK modeli sosyal ağın bağlantı yapısını içermediğinden şüpheli ağdaki kişiler arasındaki ilişkiler hakkında bilgi vermemektedir.

Bu çalışmada, benzer konu dağılımlarına göre yazar-alıcı (mesaj gönderen-mesaj alan) çiftlerini bir topluluğa tahsis eden ve ağdaki kullanıcılara ait bağlantı bilgisi ve bağlantı içeriğini keşfederek ağ hakkında maksimum bilgi alınmasını sağlayan olasılıklı yönlü bir üretici model geliştirilmiştir. Model Yazar-Alıcı-Topluluk-Konu (YATK) modeli olarak adlandırılmıştır. TK modeline bağlantı bilgisi eklenerek geliştirilmiştir. Böylece ortak ilgi alanlarına sahip toplulukların elde edilmesi yanında, topluluk altında sıralanmış üyeler arasındaki ilişkinin derecesi de gözlenebilmektedir. Geliştirdiğimiz topluluk tanıma yöntemi, şüpheli ağların analizi konusunda yeni bir pencere açmıştır.

Tezin 2. bölümünde şüpheli ağının temel alındığı çizge teorisinden, suçlu ve şüpheli ağlarda kullanılan yaklaşım modellerinden, ağın analizi için kullanılan

yöntemlerden, konu özetleme modellerinden, konu modellerinin analizinde kullanılan yöntemlerden ve topluluk tanıma modellerinden bahsedilmiştir. 3. bölümde geliştirilen YATK modeli açıklanmıştır. 4. bölümde deneylere ve karşılaştırmalara yer verilmiştir. Son olarak 5. bölümde sonuçlar ve gelecek çalışmalara değinilmiştir.

## 2. TEMEL BİLGİLER VE İLGİLİ ÇALIŞMALAR

Emniyet ve istihbarat birimleri suçlu yada şüpheli örgütleri modellemek için ağ yapılarını kullanmaktadır. Böylece örgütlerinin yapısını, devamlılığını, suç örgütüne ait grupları ve alt grupları bu ağ yapısı içinde inceleme şansı bulmaktadırlar. Bu yapı kişilerin arkadaşlık, akrabalık ve iş gibi çeşitli sebeplerle iletişime geçtiği bir sosyal ağdır. Bu ağ, üyelerin arasındaki ilişkileri temsil eden bağlantılara ve bireyleri temsil eden düğümlere sahip bir çizge yapısıdır. Bu nedenden dolayı bu bölümde ilk olarak çizge kuramı için bazı ön açıklamalar ile kullanılan matematiksel sembollerin tanıtımı yapılacaktır. Daha sonra suçlu ve şüpheli ağ analizi, geliştirilen modelin temel aldığı Bayes teoremi, konu modelleri ve topluluk tanıma modelleri açıklanacaktır.

### 2.1. Çizge Kuramı

Çizge, bir dizi çizgi ile rastgele birbirine bağlanan noktalar kümesinin oluşturduğu basit bir veri yapısıdır. Bir  $G$  çizgesi düğüm olarak adlandırılan boş olmayan elemanlar kümesi ve kenar olarak adlandırılan düğüm çiftlerini birbirine bağlayan bir liste içerir.  $G$  çizgesinin düğüm kümesi  $V(G)$  ve kenar listesi  $E(G)$  olarak gösterilir. Bir yönsüz  $G$  çizgesi

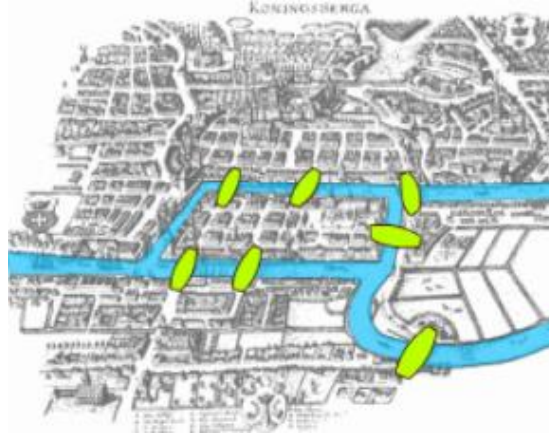
- i) boş olmayan sonlu bir  $V$  düğümler kümesinden
- ii) sonlu bir  $E$  ayrıtlar kümesinden
- iii) ve bir  $\delta: E \rightarrow P(V)$  fonksiyonundan

oluşur.

Çizgeler bir çok çeşitte ağ yapısı için model olarak kullanılmaktadır. Bu ağ yapıları için düğümler insan, bilgisayar veya şehir gibi ağın elemanını, kenarlar ise bu ağ elemanları arasındaki ilişkileri temsil etmektedir.

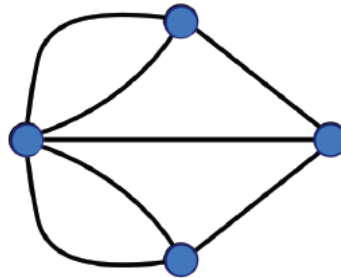
### 2.1.1. Çizgenin Tarihi

Leonhard Euler, bir çalışmasında Königsberg'in yedi köprüsü problemini çözebilen bir teori öne sürmüştür [16]. Königsberg problemi, Şekil 2.1'de görüldüğü gibi Königsberg kasabasının içinden geçen Pregel nehrinin ortasında nehrin kıyılarına ve birbirine köprüler ile bağlı iki ada ile ilgilidir. Kıyıların veya adaların birinden yola çıkan bir kasaba sakini için tüm bu köprülerden bir kere geçen ve sonunda başladığı yere geri dönmesini sağlayan bir yol bulunulması amaçlanmıştır.



Şekil 2.1: Königsberg'in yedi köprüsü.

Euler bu problemin çözümü için nehir kıyılarını ve adaları düğümlerle, köprüleri ayrıtlarla ifade ederek bir çizge yapısı kullanmıştır ve bu çözüm çizge kuramının temelini oluşturmuştur. Bu çizge yapısı Şekil 2.2'de gösterilmektedir.



Şekil 2.2: Königsberg probleminin çizge yapısı ile gösterimi.

Euler'in incelemeleri böyle bir gezintinin mümkün olmadığını kanıtlamıştır. Bu tür dolaşmayı mümkün kılmak için çizgelerin bazı özelliklere sahip olması gerekmektedir. Çizgenin sahip olması gereken özellikler [17]

*“Birleşik bir çizgenin bütün elemanlarını bir ve yalnız bir kez kullanarak dolaşmak için o çizgenin tek dereceli düğümlerinin sayısı, eğer varsa, iki olmalıdır. Tek dereceli düğümler, dolaşmanın başlangıç ve bitiş düğümleridir. Çizgede böyle düğümler yoksa dolaşmaya herhangi bir düğümden başlanılabilmektedir.”*

ifadesi ile açıklanmıştır.

Bir düğüm, başlangıç ya da bitiş düğümü değilse o düğüme gelen kişinin turu tamamlayabilmek için oradan ayrılması gerekecektir. Dolayısıyla bu tip düğümler çift dereceleri olmalıdır. Oysa tek dereceli bir düğüme, örneğin,  $D$  düğümüne ikinci kez gelen bir kişi çıkış yolu bulamayacaktır. Dolayısıyla bu düğüm ya gezintinin bitiş düğümü olmalıdır ya da başlangıç düğümü olarak seçilmelidir. Böylece ikinci gelişte çıkış yolu bulunabilir. Buna göre tek dereceli düğüm sayısı ikiden fazlaysa gezinti tamamlanamayacaktır.

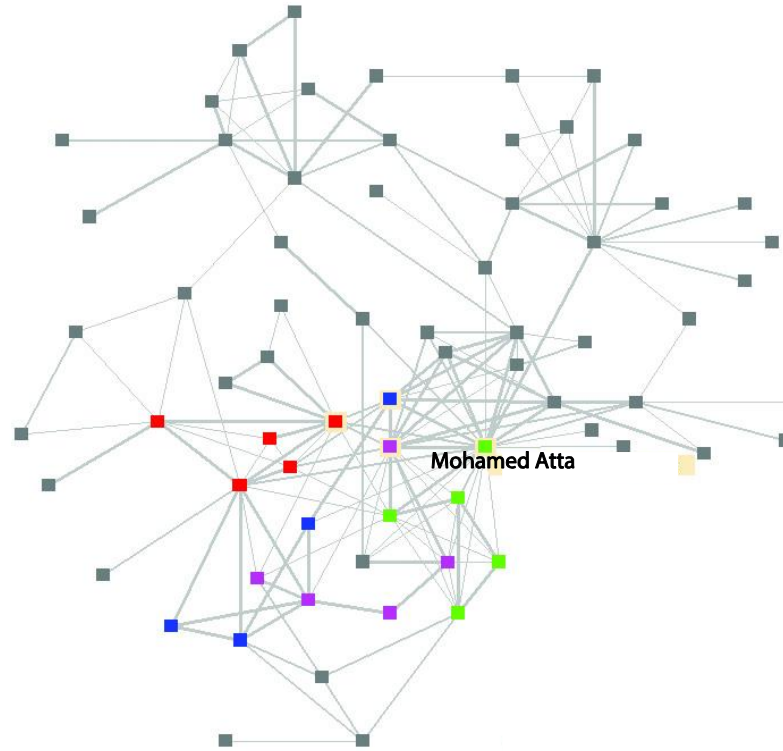
### 2.1.2. Çizge Özellikleri

Bir çizgeye ait düğüm ve kenarların karakterleri bu çizgenin özelliklerini belirlemektedir. Bir çizgede bulunan kenarların yön bildirmemesi durumunda bu çizgelere yönsüz çizge, kenarların yön belirttiği çizgelere ise yönlü çizge denilir. Yönsüz çizgelerde iki düğüm arasında bulunan kenar, her iki yönlü de hareket edilebileceğini ifade eder. İki düğüm arasındaki kenarlara bir değer verilmemiş ise bu tür çizgelere ağırlıksız çizge denir. Bir çizgedeki bütün düğümleri diğer bütün düğümlere bağlayan bir yol bulunuyorsa bu çizgelere bağlı çizge denir. Bu çalışmada kullanılan çizgeler yönlü ve ağırlıklı çizgelerdir. Düğümleri birbirine bağlayan kenarlar belli değerlere sahiptir. Ayrıca çizge bağlı bir çizge yapısında değildir. Hiç bir düğüme bağlı olmayan düğümler olabilir. Düğümler arasındaki ilişkiyi bilmek çizgenin analizi için önemlidir. Bu sebep ile, bir düğümün hangi düğümlerle komşu olduğunu gösteren bir matris tutulur.

Şehirlerin modellendiği bir çizgede şehirler arasındaki mesafe iki şehir arasındaki kenar değerine tahsis edilir. Sosyal ağların kullanıldığı durumlarda ise iki kişi arasındaki ilişkinin kuvveti kenarın ağırlığını oluşturur. Komşuluk matrislerin kullanımı bu tür ağların analizi için çok önemlidir.

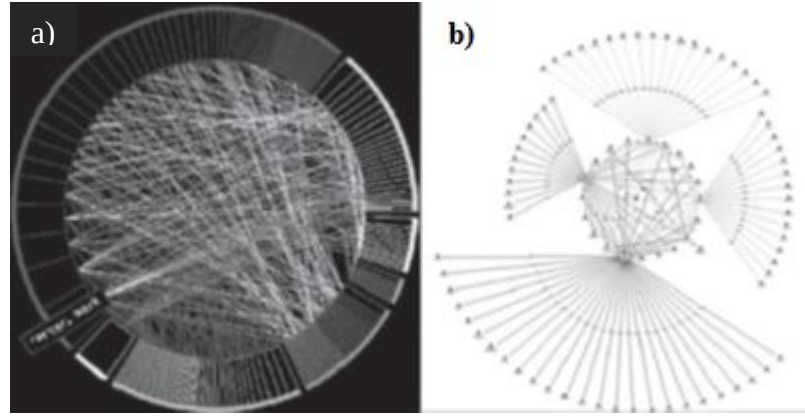
## 2.2. Suçlu ve Şüpheli Ağları

Suçlu ağı, kişilerin arkadaşlık, akrabalık ve iş gibi çeşitli sebepler ile iletişime geçtiği sosyal ağıdır. Bir sosyal ağ, üyelerin arasındaki ilişkileri temsil eden bağlantılara ve bireyleri temsil eden düğümlere sahip bir çizge yapısı olarak görülür. Bir sosyal ağın yapısı, düğüm, bağlantı, grup ve bütün ağ olarak tanımlanmıştır. Suçlu ağı oluşturma yaklaşımları genellikle üç farklı başlık altında anlatılmaktadır [18]. İlk kuşak yaklaşımı Anacpapa Grafik [19] temsil eder. Bir araştırmacı suçlular arasındaki bağlantıları tanımlamak için verileri araştırır ve suçlular arasındaki bağlantıları tanımlayarak ilişkisel bir matris oluşturur. Bu matris kullanılarak bir tablo çizilir. Diğer kişilerle bağlantısı en fazla olan suçlu genellikle tablonun merkezine yerleşir. Örnek olarak Krebs'in ağı verilebilir [20]. Krebs teröristlerin birbirleri ile ilişkilerine dair gazetelerden topladığı bilgilerle bir ilişkiler matrisi modelleyerek Şekil 2.3'te görüldüğü şekliyle bir suçlu ağı oluşturmuştur. Bu ağ incelendiğinde, 19 uçak korsanı arasında en fazla bağlantıya sahip olan ve bu ağın lideri rolünü üstlenen kişinin Mohamed Atta olduğu keşfedilmiştir. Mohammed Atta Dünya Ticaret Merkezi'ne çarpan ilk uçağın pilotudur.



Şekil 2.3: Krebs'in 19 uçak korsanına ait terörist ağı.

Suçlu ağı analizi için kullanılan ilk kuşak yaklaşım, suç araştırmada yardımcı bir yöntem olmasına karşın, suçlu ağı büyüdükçe oldukça yavaş ve yetersiz kalmaktadır. İkinci kuşak yaklaşım araçları, otomatik olarak bir ağın grafiksel görünümünü üretir. Günümüzde de kullanılan Netmap [21] ve Xanalyst [22] en popüler olan modellerdir.

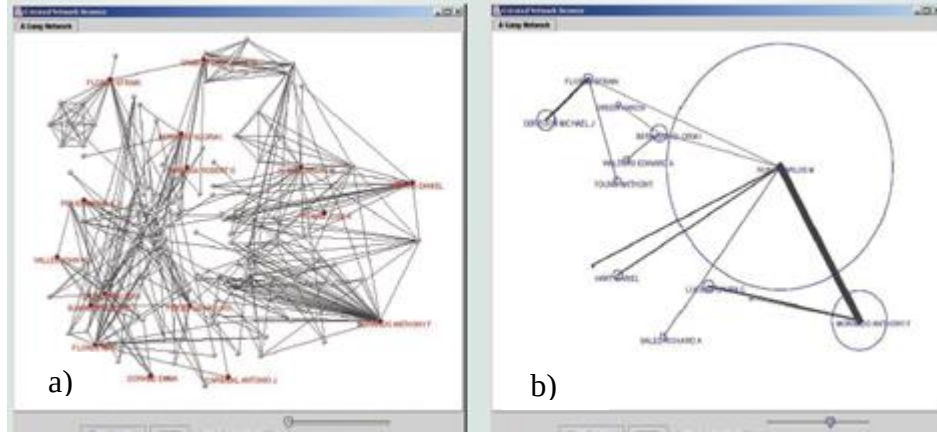


Şekil 2.4: a) Netmap’a ait ağ görüntüsü, b) Xanalyst’e ait ağ görüntüsü.

Netmap çeşitli tipte bilgileri bir çember üzerinde konumlandırır ve bilgiler arasındaki bağlantıları göstermek için veriler arasına düz çizgi çizer. Bu çizgilerin kalınlıkları temsil ettikleri bağlantının kuvvetini gösterir. Araştırmacı, bu bağlantılara bakarak kendisi için gerekli olan örnekleri yakalar. Netmap kara para aklamayı engellemek için finansal hareketlere ait örnekleri analiz etmek için kullanılmıştır. Xanalyst ise verilen bir suçlunun ismine göre veri tabanından arama yaparak o suçlu ile ilişkisi olan kişilerin isimlerini ve bağlantılarını görüntüler.

Üçüncü kuşak yaklaşımda ağı yapısı ile ilgili bilgiler elde edilmesi amaçlanmıştır. Ağı yapısal karakteristikleri keşfedilerek, üyeler, gruplar, gruplar arasındaki bağlantılar ve bütün ağ yapısı hakkındaki bilgiler, geniş bir veri işlenerek elde edilir. Yapısal ağ analiz yaklaşımı için sosyal ağ analiz dinamikleri kullanılmaktadır. Şekil 2.5’te bir suçlu ağına ait ekran görüntüsü verilmektedir. Polis departmanlarından elde edilen suçlular ile ilgili bilgiler kullanılarak suçlular arasındaki bağlantılar görüntülenmiştir. Fakat 164 tane şüpheli ve çok sayıda bağlantı olduğu için bu ağı bu şekilde analizi zordur. Bu nedenle her bir şüpheliye ait alt gruplarında görüntüsü elde edilmektedir. Bu çizgedeki düğümler bir çember olarak gösterilmektedir ve her bir çember bir şüpheliyi yada suçluyu temsil eder. Bu çemberlerin boyutu o düğümün kenar sayısı ile orantılıdır. Kenar sayısı bilgisi

genellikle suçlu ağlarında o düğümün aktifliği diğer bir deyişle o alt grup içerisindeki önemi hakkında bilgi vermektedir. Çemberin boyutu büyüdüğüçe o şüphelinin ağ içerisinde aktif bir kişi olduğu düşünülür. Üçüncü kuşak yaklaşım ağ hakkında bilgi verebildiği için suçlu ağlarının analizinde tercih edilen yöntem olmuştur.



Şekil 2.5: Yapısal ağ analiz yaklaşımı, a) Tüm ağ görüntüsü, b) Alt gruplar.

### 2.2.1. Suçlu ve Şüpheli Ağları Analizi

Emniyet ve istihbarat birimleri suç örgütlerini modellemek için ağları kullanmaktadır. Bunun nedeni, suç örgütlerinin yapısını, devamlılığını, grup ve alt gruplarını bu ağ yapısı içinde incelemektir. Uyuşturucu kaçakçılığı, çete ile ilişkili olaylar, dolandırıcılık, silahlı soygunlar, terör gibi suçlar organize suçlar olarak adlandırılır. Suçlular şirketler gibi kendi yapılarını oluştururlar. Bu yapının içinde farklı görevlerden sorumlu farklı takımlar ve bireyler vardır. Ağ yapılarını, işlemlerini ve bireysel rollerini anlamak suçları engellemek için yeni stratejiler geliştirmeye yardım eder. Fakat suçlu ağı analizi çeşitli zorluklara sahiptir. Bunlardan en önemlisi suçluların oluşturduğu ağın statik olmayışıdır. Sürekli bir değişim vardır. Bu ağa yeni üyeler katılabilir ya da üyelerden ayrılanlar olabilir. Suçluların sorumlu olduğu görevler değişebilir, güçlenebilir ya da güç kaybına uğrayabilir. Üyeler arasındaki ilişkiler kopabilir ya da güçlenebilir. Gruplar birleşebilir ya da bir grup birkaç gruba bölünebilir. Bütün ağdaki hiyerarşik yapı bozulabilir ve yeniden yapılanabilir. Bu değişiklikler suçlu ağı analizini zorlaştırır. Bu tür değişken özellikler ağın dinamik özelliğini oluşturur. Bu özellikler araştırmacılara inceledikleri ağ ile ilgili çok önemli bilgiler verebilir. Suçlu ağının

dinamik özelliklerini yakalamak için bazı tekniklere ve metodolojilere gerek duyulmaktadır. Bu teknikler genellikle sosyal ağ analizinde kullanılan tekniklerdir. Bu dinamik ağları analiz etme yaklaşımları tanımlayıcı, istatistiksel ve benzetim olarak ayrılmaktadır [23], [24].

Tanımlayıcı yöntemin amacı, sosyal ağlardaki yapısal değişiklikleri fark etmek ve deneysel veriler üzerinde çalışarak teorileri ne kadar desteklediğini test etmektir. Ağın yapısal özellikleri çeşitli ölçüm ve karşılaştırmalarla ölçülür. Bir ağın düğüm, bağlantı ve gruplardan oluşmasından dolayı, ölçüm ve karşılaştırmalar düğüm, bağlantı ve grup seviyesinde olacaktır.

Düğüm seviyesinde ölçümlerin amacı, suçluların özelliklerine göre bir sınıflama yapmak ve değişiklikleri tespit etmektir. Bu tespit için kullanılan en popüler ölçümler, Freeman [25] tarafından üç başlıkta kategorize edilmiştir. Bunlar derece, aralık ve kapalılık değeridir.

Derece, bir düğümün ne kadar aktif olduğunu ölçmede kullanılır. Bir  $k$  düğümünün sahip olduğu bağlantıların sayısı düğümün derecesini gösterir. Eğer bir düğümün derecesi yüksek ise, bu o grubun ya da ağın lideri olduğunu gösterebilir. Bazen lider kendini gizleme yoluna gittiği için bağlantı sayısı en az olan kişide yani düğüm derecesi en düşük olan kişi de lider olabilir. Böyle bir durumda düğüm derecesi en yüksek olan kişi gruptaki liderden sonraki ikinci eleman olabilir. Suç ve suçlu profiline göre değişik yaklaşımlarda bulunulabilir.  $n$  sayıda düğümden oluşan bir suçlu ağında  $k$  düğümünün aktifliği eşitlik (2.1)'de verilen  $C_D(k)_{norm}$ 'un hesaplanması ile bulunur.

$$C_D(k)_{norm} = \frac{\sum_{i=1}^n a(i, k)}{n-1}, \quad i \neq k \quad (2.1)$$

Bir  $k$  düğümü için aralık değeri,  $k$  düğümden geçen diğer iki düğüm arasındaki en kısa yolların sayısıdır. Bir kişi için yüksek aralık değeri, o kişinin örneğin, bir uyuşturucu şebekesinde uyuşturucuyu dağıtan bir aracı olduğunu gösterebilir.  $n$  sayıda düğümden oluşan bir suçlu ağında  $g_{ij}(k)$ ,  $k$  düğümünden geçen  $i$  ve  $j$  düğümleri arasında en kısa yolun olup olmadığını gösterir. Eşitlik (2.2) ağ için aralık değerinin hesabını vermektedir.

$$C_B(k)_{norm} = \frac{\sum_i^n \sum_j^n g_{ij}(k)}{n-1}, \quad i \neq k \quad (2.2)$$

Kapalılık değerine ulaşmak için, bir  $k$  düğümünün, ağdaki diğer düğümlerden ne kadar uzak olduğunu ölçmek gerekir. Bu değer bir kişinin ağdaki diğer üyelerle ne kadar kolaylıkta erişebildiğini gösterir.  $l(i, k)$ ,  $i$  ve  $k$  düğümlerini bağlayan en kısa yolların uzaklığını verir.  $C_c min$ , en kısa yolun minimum uzaklık değerini,  $C_c max$  ise en kısa yolun maksimum uzaklık değerini verir. Kapalılık değeri eşitlik (2.3)'te verilmiştir.

$$C_C(k)_{norm} = \frac{\sum_{i=1}^n l(i, k) - C_c min}{C_c max - C_c min}, \quad i \neq k \quad (2.3)$$

Bir ağdaki suçlular birbirleri ile olan bağlantılarına göre gruplandırılabilir. Oluşan bu grupların yapısal özelliklerini analiz etmek için en popüler olan üç ölçüm kullanılmaktadır. Bunlar: bağlantı yoğunluğu, grup bağlılığı ve grup dengesidir. Bu ölçümler grup seviyesinde ölçümlerdir.

Bütün grup üyeleri arasında var olan maksimum bağlantı sayısının oranı bağlantı yoğunluk değerini verir. Eğer bu oran 1 ise, bu grubun üyelerinin her birinin grubun diğer üyeleri ile iletişimde olduğu sonucuna varılır. Bir grup içindeki bir suçluya erişildiğinde diğer suçlulara erişim kolay olacaktır.  $n$  sayıda düğüme sahip olan bir grup için yoğunluk değeri eşitlik (2.4)'de verilmektedir.

$$Yoğunluk = \frac{\sum_i^n \sum_j^n a(i, j)}{n(n-1)} \quad (2.4)$$

Grup bağlılığı, bir grubun içindeki üyelerin birbirleri ile olan bağlantılarının, diğer bir gruba kıyasla ne kadar güçlü olduğunu ölçer. Bağlılık değeri 1 olursa, bu grubun bağlı bir grup olduğu sonucuna varılır. Bağlı gruplarda, üyelerin gruba sadakati daha yüksektir ve genellikle üyeler gruplarından ayrılıp başka bir gruba geçmezler.  $N$  sayıda düğüme sahip olan bir suçlu ağı ve  $n$  sayıda düğüme sahip

olan bir grup için bağıllık değeri eşitlik (2.5)'te verilmektedir.  $w(i, j)$ ,  $i$  ve  $j$  düğümleri arasındaki bağlantının ağırlığıdır.

$$Bağıllık = \frac{\sum_{i \in G} \sum_{j \in G} w(i, j)}{n(n-1)} \quad (2.5)$$

$$= \frac{\sum_{i \in G} \sum_{j \in G} w(i, j)}{N(N-n)}$$

Grup dengesi, üyelerin bir grupta ne kadar kaldığını gösteren bir değerdir. Grubun iki farklı zamandaki üyelikleri temel alınarak grup dengesi hesaplanır. Bir  $G$  grubunun eşitlik (2.6)'da görüldüğü üzere  $t_1$  ve  $t_2$  zamanı kıyaslanarak elde edilir.

$$Denge = \frac{|G_{t_1} \cap G_{t_2}|}{|G_{t_1} \cup G_{t_2}|} \quad (2.6)$$

İstatistiki yöntem, ağdaki değişimleri tanımlamanın ve keşfetmenin ötesinde, bu değişikliklerin neden olduğunu açıklayan bir yaklaşım öne sürmektedir. Ağ ile ilgili yapısal değişikliklerin karşılık, geçişlilik ve denge gibi süreçlerin sonucu olduğunu varsayar. Bu yöntemde, düğümler arasındaki bağlantılar pozitif, negatif ya da nötr olabilir. Ayrıca farklı zamanlarda farklı durumlarda olabilecek şekilde modellenebilir. Ayırık ve sürekli Markov modeller [26] istatistiksel yöntemlerde kullanılmaktadır. Markov yönteminde  $n \times n$ 'lik bir suç nedenselliği matrisi oluşturulur [27]. Bu matris suçun birincil nedeni, ikincil nedeni ve sonuç değerleri ile ilişkilidir.

Leenders bir çalışmada çocukların cinsiyet benzerliği, karşılık ve akrabalık parametrelerine göre arkadaş seçiminde hangi parametreyi önemseydiğine dair bir araştırma yapmış ve sürekli zamanlı Markov model kullanmıştır [28]. Leenders'ın araştırmasının sonucunda, cinsiyet benzerliğinin çocukların arkadaş seçiminde en önemli kriter olduğu görülmüştür.

Benzetim yöntemi, ağ üyelerinin ağ içindeki davranışlarını modelleyerek ağın bir yapıdan bir başka yapıya dönüşümüne karar verir. Ağ içindeki bireylerin seçimlerine göre ilişkilerin kurulması ya da bozulmasına dair değişikliklerin ağın yapısını nasıl etkilediğini göstermek için bir benzetim kullanmıştır [29]. Bu çalışmaya göre, bir ilişki kişinin amacı, yararı ve ilişki maliyetine göre devam eder.

## 2.3. Bayes Olasılığı ve Konu Modelleri

### 2.3.1. Bayes Olasılığı

Bayes olasılık terimi Thomas Bayes'in 1763 yılında yayınladığı bir kitap ile ortaya atılmıştır [30]. Fakat Pierre-Simon Laplace'ın, bu teorimi tıbbi istatistik ve güvenilirlik tahminleri gibi problemler için kullandığı zaman en genel halini kazanmıştır. Bu teorem bir rassal değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi gösterir. Ancak Bayes teoremi objektif ve sübjektif olarak adlandırılan iki farklı görüşü de içerisine almaktadır. Felsefi temelde olasılık değerlerinin nesnel bir özellik değil, gözlemcinin meydana çıkardığı sübjektif bir değer olarak kabul eden sübjektivist olasılık düşünürlerine göre Bayes teoremi, yeni kanıtlar ışığında olasılık değeri hakkındaki sübjektif inanışların güncelleştirilip değiştirilmesini sağlayan temel bir gereçtir.

Olasılık teorisi içinde incelenen bir  $X$  olayının bilindiği  $Y$  durumunda  $Y$  olayı için olasılık değeri ile  $Y$  olayı bilindiği durumdaki  $X$  olayı için olasılık değeri birbirinden farklıdır. Fakat bu iki birbirine ters koşulluluk arasında bir ilişki vardır ve bu ilişkiye Bayes Teoremi denilmektedir. Formel bir teorem olarak Bayes teoremi, olasılık kavramını inceleyen her çeşit değişik felsefi temel fikre bağlı olan her türlü istatistikçi tarafından kabul edilir. Ancak olasılığı objektif bir değer olarak gören ve relatif çokluluk olarak tayin eden frekansçı ekole bağlı olan istatistikçiler ile sübjektivist ekoline bağlı olan istatistikçiler arasında bu teoremin pratikte nasıl kullanılabileceği hakkında büyük bir fikir ayrılığı bulunmaktadır. Frekansçı ekolüne dahil olanlar olasılık değerlerini rastgele olaylarda meydana çıkma çoğulluğuna göre veya ana kütlelerin alt setlerinin tam ana kütleyle orantısı olarak saptanması gerektiğini kabul etmektedirler. Yeni kanıtlar karşısında olasılık değerinin değişme imkânı yoktur. Bu nedenle frekansçı ekol için Bayes teoremi sadece koşulluluklar arasında ilişkiyi gösterir ve bunun pratikte kullanılma gücü küçüktür. Hâlbuki sübjektivist ekolüne göre olasılık gözlemcinin sübjektif belirsizlik ifadesidir. Bu nedenle olasılık değeri sübjektif olup, yeni kanıtlar geldikçe değiştirilebileceğine inanmakta ve böylece Bayes teoremini istatistik bir incelemenin temel taşı saymaktadırlar.

Bayes teoremi bir rastlantısal süreç sırasında ortaya çıkan bir rastgele olay  $X$  ile bir diğer rastgele olay  $Y$  için koşullu olasılıkları ve marjinal olasılıkları

arasındaki ilişkidir ve eşitlik (2.7)'de gösterilmektedir.  $p(D)$ , bütün olası varsayımlar altında  $D$  olma durumunun önsel olasılığıdır.  $H_i$  varsayım kümesi verildiğinde,  $p(D)$  eşitlik (2.8)'de yazıldığı gibi hesaplanmaktadır.

$$p(H | D) = \frac{p(D | H)p(H)}{p(D)} \quad (2.7)$$

$$p(D) = \sum_i p(D | H_i) = \sum_i p(D | H_i) p(H_i) \quad (2.8)$$

$X$  ve  $Y$  rastgele değişkenleri düşünüldüğünde,  $p(X,Y)$  her iki rastgele değişkenin ortak dağılımını gösterir. Rastgele bir  $X$  değişkenin marjinal dağılımı eşitlik (2.9) kullanılarak ortak dağılımdan elde edilir.

$$p(X) = \int_Y p(X,Y) dY \quad (2.9)$$

$Y$  olayı verildiğinde  $X$  'in koşullu olasılığı,  $Y$  olayı gerçekleştiğinde  $X$  olayının olasılığıdır. Olasılık çarpım kuralı eşitlik (2.10)'da gösterilmektedir.

$$p(X,Y) = p(Y | X)p(X) \quad (2.10)$$

### 2.3.2. Frekans Olasılığı

Frekans olasılığı kavramı, kendini ölçülebilirliğe kısıtlama fikrini temel alır. Bayes olasılığındaki gibi rassal değerler yoktur. Sadece her zaman olan yada olmayan rassal olaylar üzerinde sınırlanır. Veri gözlenen şeyin ne olduğuna göre yorumlanır ve ölçülebilir. Bir sahte bozuk para düşünüldüğünde, bozuk para atıldığı anda paranın bir yüzünün gelme olasılığını tahmin etme olayı bu model için uygun bir örnektir. Fakat bu oyunu oynayan iki kişiden biri hile yapar ise, bunun olasılığı nasıl etkileyeceğini anlamak mümkün olmaz. Aksine bu varsayım başlangıçta kabul ya da ret edilir. Frekans olasılığı bir olayın nispi frekansı ile ilgilenir. Uzun vadede, bütün olayların ve iddiaların tamamı verildiğinde, bu frekans olayın olasılığına eşittir.

Rastgele bir deneyde,  $n$  adet denemede, bir  $e$  olayı  $n_e$  sayıda görüldüğü varsayıldığında, nispi yani göreceli frekans  $n_{relative}$  eşitlik (2.11)'de gösterilmektedir.

$$n_{relative} = \frac{n_e}{n} \approx p(e) \quad (2.11)$$

Frekans olasılığı için örnek vermek gerekilirse, verileri analiz etmek için kullanılan yöntemlerden en basit yöntem kelime torbası yöntemidir. Her doküman sahibi için, yazarın dokümanlarda kullandığı kelimeler bir sayılır ve frekans vektörü oluşturularak kaydedilir. Bundan sonra, vektör sözlük boyutuna eşit bir sayısal bölünür. Bu vektör bir olasılık dağılımı olarak tanımlanamasa da, bilgi yarıcapı ve eğriltme diverjans uzaklık ölçümleri ile iki vektör yani yazar arasındaki mesafeyi hesaplamak için kullanılır. Kelime torbası modeli frekans vektörlerini kullandığı için iki vektör arasındaki benzerlikleri ölçmek için benzerlik yöntemleri kullanılabilir. Eğer iki yazarın dokümanları karşılaştırıldığında, ortak ilgi alanları var ise dokümanlar arasındaki benzerlik değeri yüksek çıkacaktır. Bu çalışmada, verilerin oluşturulmasında bu yaklaşım kullanılmaktadır.

### 2.3.3. Üretken Modeller

Bir olasılıklı model içerisindeki gizil değişkenler, doğrudan gözlenemeyen gerçek değerleri gizli olan rassal değişkenlerdir. Gizil değişkenleri doğrudan gözlenebilen ve ölçülebilen diğer değişkenlere bağlayan bir istatistiksel model kullanılır. Bu model ışığında dolaylı olarak anlaşılabilirler. Kullanılan bu modele gizli değişken modeli denir. Bu modelin amacı gözlenebilir değişkenler üzerindeki dağılımı etkileyen gizli değişkenlere ait bir dağılım bulunarak gözlenebilir değişkenleri açıklamaktır. Bu sebeple, gözlenen değerler ile gizli değerler üzerinden bir bileşik olasılık dağılımı tanımlanır ve Bayes teoremi kullanılarak gözlenen değerler üzerinden uygun dağılıma karar verilir. Gizli değişken modelleme giderek istatistiğin ayrılmaz bir parçası haline gelmiştir ve şu anda biyometrik genetiği, veri analizi ve psikometrik ölçüm gibi farklı alanlarda çok sayıda uygulama için kullanılmaktadır.

Bu çalışmada kullanılan gizil model için matematiksel gösterim ve terim bilgisi aşağıdaki gibidir.

- Verinin en temel birimi kelimedir. *Kelime*,  $\{1,...,V\}$  kümesi içinde indekslenmiş bir sözlüğün küçük bir parçasıdır. Sözlükte  $v$ . kelime olarak gösterilen kelime  $w_v$  olarak gösterilir.
- Bir *doküman*,  $N$  sayıda kelimeden oluşur ve  $\mathbf{w} = (w_1, w_2, \dots, w_N)$  vektörü olarak gösterilir.
- Bir *yapı*,  $M$  sayıda dokümandan oluşur ve  $D = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$  vektörü olarak gösterilir.

Gizil değişken modeli, mevcut verileri analiz ederek gözlenemeyen değişkenleri anlamaya çalışır. Büyük bir veri yapısına ait dokümanlar içerisinde gizli bir yapıyı diğer bir deyişle, aranan ama önceden gözlenemeyen bir bilgiyi bulmak için üretken bir model kullanılır. Kısacası mevcut verileri analiz ederek büyük bir veri yapısının özetini çıkartan ve basit olasılık kurallarını temel alarak çalışan bir modeldir. Bu modelin amacı, gözlenen yapıyı açıklayan en iyi gizil değişken kümesini bulmaktır. Üretken model, yapılandırılmış temsil ve istatistiksel öğrenme tekniklerini birleştirmektedir. Yapı ve doğal dil analizi için tüm teknikler göz önüne alındığında, verinin boyutunu azaltmak için kullanılan üretken modelin diğer metotlara göre daha avantajlı olduğu görülmektedir. İnternet arama motorlarında kullanılan TF-TDF yöntemi [31] en basit metodolojidir. Bu yöntem, bir yapı içerisindeki her doküman için bir kelimenin ne kadar önemli olduğunu yansıtan sayısal istatistikler verir. Bir yapı içerisindeki her doküman için bir sayı hesaplar. Bu sayı bir kelimenin ne kadar sıklıkta kullanıldığı bilgisini verir. Uygun bir normalleştirmeden sonra, bu sıklık sayısı, tüm yapı içerisinde bir kelimenin ne kadar sıklıkta kullanıldığı bilgisini ölçen ters doküman frekans sayısı ile karşılaştırılır. Bu yaklaşım küçük bir miktarda veri boyutunu azaltmayı sağlamaktadır. Bu problemi çözmek için bir çok yöntem geliştirilmiştir. Eş anlamlılık ve anlamlılık gibi bazı dilsel bilgileri yakalayan GAİ [32] bu modellerden birisidir. Hofmann tarafından ortaya konulan olasılıklı GAİ (OGAİ) modeli [33], bu çalışmalar için en önemli adımdır. Bu yaklaşım, bir doküman içerisindeki her kelimeyi bir karışma modelinden bir örnek alarak modeller. Karışma bileşenleri ise konuları temsil eden çok terimli

rassal deęişkenlerdir. Bu alıřmada, her kelime bir konudan üretilir ve bir doküman içindeki farklı kelimeler, farklı konulardan üretilir. Bu yaklaşımda, her doküman bir sayılar listesi ile dięer bir deyişle konular için karışırma oranı ile temsil edilir. Bu yöntem için olasılıklı üretken bir model yoktur. Bu durum, model içerisindeki parametre sayısının, yapının büyüklüğüne baęlı olarak lineer şekilde artmasına ve bu yapının dışında bir doküman için bir olasılık tahsis etme probleminin özölmemesine sebep verir.

GAİ ve OGAİ yöntemlerinin her ikisi de kelime torbası yöntemini kullanmaktadır. Çok büyük boyutlu uzaylarda metin sınıflama, özet bilgi ıkarımı gibi konular üzerine alışıldığında kelimeler kümelenerek metinler daha az boyutlu uzaylarda temsil edilmektedir. Böylece daha küçük boyutlu bir modelleme saęlanarak daha başarılı sonuçlar elde edilmektedir. İşlem zamanının kısalması ve hafıza ihtiyacının azalması sebebi ile kelime torbası yöntemi tercih edilmektedir. Her doküman kelimelerden oluşur. Öncelikle dokümandaki kelimeler ve gereksiz kelimelerin ayrıştırılması gerekir. Bir kelimenin dokümanda kullanılma sayısı bilgi olarak tutulur. Kelimelerin hangi sırada kullanıldığının literatürdeki haliyle deęişebilirliğin bir önemi yoktur. Her kelime sözlükte indeks deęerine sahiptir. Her doküman için indeksler ve kelimelerin dokümanda kaç kez getięi sayı bilgisine ait vektörler tutulmaktadır. Bu yöntemin kelimelerin sırasını göz ardı etmesi durumu GDT modelinin oluşturulmasına önderlik etmiştir.

#### 2.3.4. Markov Zinciri Monte Carlo Yöntemi

oęu olasılık modellerinde kesin ıkarımlar yapmak ve sonuçlar vermek çok zordur. Bunun için yaklaşık deęerler bulmak ve bu deęerleri bulan eşitli yöntemler kullanmak gerekmektedir. Bunu yapmak için kullanılan en bilinen yöntem Monte Carlo yöntemidir. Bu yöntem sayısal örneklemeye dayalı bir yaklaşım ıkarım yöntemidir.  $z$ 'nin ayırık ya da sürekli bir deęişken olduęu bir  $p(z)$  olasılık dağılımı ile ilgili bir  $f(z)$  fonksiyon beklentisi bulur. Beklenti deęerlendirilmesi eşitlik (2.12)'de verilmektedir.

$$E(f) = \int f(z) p(z) dz \quad (2.12)$$

Fakat bu hesaplama çok karmaşık olduğu için Monte Carlo Metodu ile  $p(z)$  olasılık dağılımından bağımsız çizilen  $l=1,\dots,L$  olduğu  $z_l$  örnekler kümesi elde edilir ve beklenti  $f$  sonlu bir toplam kullanılarak bir yaklaşım olarak verilir.

$$\bar{f} = \frac{1}{L} \sum_{l=1}^L f(z_l) \quad (2.13)$$

$z_l$ ,  $p(z)$  olasılık dağılımından çizildiği için  $E[f] = E[\bar{f}]$  olmaktadır. Ayrıca yüksek bir doğruluğa erişmek için bazen küçük bir örnek miktarı yeterli olabilmektedir. Bu durum genel doğruluğun  $z$ 'nin boyutuna bağlı olmadığını gösterir.  $f$  hesabı eşitlik (2.13)'te verilmektedir.

Matematikte, Markov Zinciri, Markov özelliğine sahip bir rastgele süreçtir. Markov özelliğine sahip olmak, mevcut durum verildiğinde, gelecek durumların geçmiş durumlardan bağımsız olması anlamına gelir. Bir başka deyişle, mevcut durumun açıklaması, sürecin gelecekteki evrimini etkileyebilecek tüm bilgiyi kapsar. Gelecek durumlara olasılıklı bir süreçle ulaşılabilecektir. Her bir anda sistem belirli bir olasılık dağılımına bağlı olarak kendi durumundan başka bir duruma geçebilir yahut aynı durumda kalabilir. Çeşitli durum değişimleriyle ilişkili olasılıklar da geçiş olasılıkları olarak adlandırılır.

Markov zinciri Markov özelliğini taşıyan, yani mevcut durum verilmiş olursa geçmiş ve gelecek durumların bağımsız olduğu  $m \in \{1, \dots, M-1\}$  olduğu yerde,  $z_1, z_2, \dots, z_M$  ile ifade edilen, bir seri rassal değişkendir. Matematiksel biçimi eşitlik (2.14)'de gösterilmektedir.

$$p(z_{m+1} | z_1, \dots, z_m) = p(z_{m+1} | z_m) \quad (2.14)$$

Markov zincirini belirtirken, sonraki değişkenler için koşullu olasılıklar diğer bir deyişle geçiş olasılıkları ile beraber  $p(z_0)$  başlangıç değeri için bir olasılık dağılımı vermek gereklidir. Dağılım eşitlik (2.15)'te gösterilmektedir.

$$T_m(z_m | z_{m+1}) = p(z_{m+1} | z_m) \quad (2.15)$$

Eğer her Markov zinciri adımında, bu dağılım sabit kalırsa, Markov zinciri değişmez hale gelecektir. Eğer

$$p^*(z) = \sum_z T(z, z) p^*(z) \quad (2.16)$$

ise  $p^*(z)$  ve  $T(z, z)$  ile birlikte Markov zinciri sabit hale gelecektir.

Belli bir süre sonra kararlı duruma gelen bir Markov zincirine denge dağılımı adı verilmektedir. Gibbs örnekleme, Markov zincirinin denge dağılımı yaklaşımı için kullanılmaktadır. Gibbs örneklemenin temeli matematiksel fizik alanına uzanmaktadır. İlk olarak çıkarımlı istatistik alanında Geman and Geman tarafından kullanılmıştır [34]. Gibbs örnekleme detaylı olarak anlatılacaktır.

Başlangıç durumunda bir Markov zinciri ve  $p(z) = \{z_1, z_2, \dots, z_M\}$  dağılımı düşünüldüğünde,  $z_i$ 'nin değeri  $z$ 'nin  $i$ 'ninci bileşeni olan  $z_i$  ve  $z_1, \dots, z_M$ 'den  $z_i$ 'nin çıkarılması ile elde edilen  $z_{\setminus i}$  olduğu durumda  $p(z_i | z_{\setminus i})$ 'den çizilen bir değer ile yer değiştirir. Bu süreç tüm değişkenler boyunca sürekli bir çevrim halinde tekrarlanır. Bu durumun dengeli bir dağılıma gittiğini göstermek için  $p(z)$ 'nin Gibbs örnekleme sürecinin her bir adımında bir değişmeyen olduğunu göstermek gerekir.

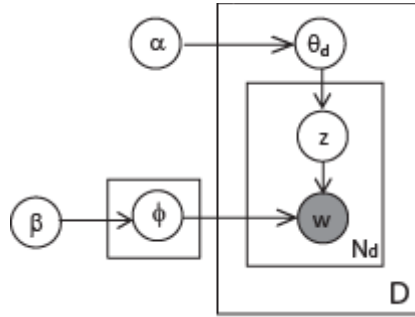
### 2.3.5. Gizil Dirichlet Tahsisi

Olasılık konu modelleri geniş miktarda arşivlerde bulunan dokümanların içeriklerini ve dokümanların sahip olduğu kelimelerin anlamlarını analiz etmek için kullanılmaktadır. GDT bir üretici olasılık konu modelidir [2]. GDT kelimelerden oluşan bir belgenin içeriğini dokümanı özetleyen başlıklar olduğunu söyler. Bu kelimeler ve başlıklar doğal bir metin içerisinde olasılıklı örneklere göre dağılmıştır. Doğal dil, insana bir konuyu anlatmak için birbirinden farklı birçok kelimeyi kullanmasına izin vermektedir. Bu kelimeler birbiri ile ilişkili yada ilişkisiz görülebilir. Bu kelimeler belli konu başlıkları altında toplanabilir. GDT modelinin temelinde yatan fikir, geniş bir belge kümesi içerisindeki her belge, belli bir konu başlığı dağılımı düşünüldüğünde bir yada birden fazla konu ile ilişkili olabilmektedir. Bir yapı içerisindeki dokümanlar içerisindeki kelimelerin sırasının bir önemi yoktur.

Bu varsayım yapı içerisindeki bir dokümana koşullu bağımsız diğer bir deyişle değiştirilebilir rassal bir değişken olarak davranılmasını sağlar. GDT modelinin ana fikri bir kelime dağılımından oluşan konular ve rassal karışımlardan oluşmuş konular olarak tasvir edilen dokümanların temelinde yatan gizil yani örtülü konu yapısını anlamaya çalışmaktır. GDT modeli, bir  $D$  yapısı içerisindeki her  $w$  dokümanı için aşağıdaki üretici süreci kullanır [2].

- Adım 1:  $N$  'i seç,  $N \sim \text{Poisson}(\epsilon)$
- Adım 2:  $\theta$  'yı seç,  $\theta \sim \text{Dir}(\alpha)$
- Adım 3:  $\varphi$  'yi seç,  $\varphi \sim \text{Dir}(\beta)$
- Adım 4:  $n$  sayıda kelimeye sahip her bir doküman için  $\mathbf{w} = (w_1, w_2, \dots, w_n)$
- Adım 4.1: Bir konu seç,  $z_n \sim \text{Multinomial}(\theta)$
- Adım 4.2: Bir kelime seç,  $w_n \sim \text{Multinomial}(\varphi_{z_n})$

Modelin plaka gösterimi Şekil 2.6'da verilmektedir.



Şekil 2.6: GDT modelinin plaka gösterimi.

Dirichlet dağılımının  $k$  ve  $j$  boyutunun sabit olduğu ve bilindiği kabul edilir. Konu değişkeni  $z$  'nin boyutu  $k$  değerini belirler. Ayrıca doküman uzunluğu  $N$ , diğer  $\theta$ ,  $\varphi$ ,  $z$  değişkenlerinden bağımsız olduğu için göz ardı edilebilmektedir.  $\theta$ ,  $k$  boyutlu Dirichlet rassal bir değişkendir ve  $(k-1)$  simpleks'in içinde değerler alır. Bu simpleksdeki olasılık yoğunluğu eşitlik (2.17)'de verilmektedir.

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1} \quad (2.17)$$

$\varphi$  'nin olasılık yoğunluğu eşitlik (2.18)'de verilmektedir.

$$p(\varphi|\beta) = \frac{\Gamma(\sum_{i=1}^j \beta_i)}{\prod_{i=1}^j \Gamma(\beta_i)} \varphi_1^{\beta_1-1} \dots \varphi_j^{\beta_j-1} \quad (2.18)$$

$\alpha$  ve  $\beta$  parametreleri, sırasıyla  $k$  boyutlu ve  $j$  boyutlu vektörlerdir.  $\Gamma(x)$  ise gamma fonksiyonudur.  $\alpha$  ve  $\beta$  parametreleri verildiğinde, konu dağılımı  $\theta$  'nın ve konu kelime dağılımının ortak olasılık dağılımı eşitlik (2.19)'da verilmektedir.  $p(z_n | \theta)$ ,  $z_n^i = 1$  olduğunda tek bir  $i$  için  $\theta_i$  'dir.

$$p(w, z, \theta, \varphi | \alpha, \beta) = p(\theta | \alpha) \prod_{j=1}^T p(\varphi_j | \beta) \prod_{n=1}^N p(z_n | \theta) p(w_n | \varphi_{z_n}) \quad (2.19)$$

Tek bir doküman için marjinal dağılım eşitlik (2.20)'de verilmektedir.

$$p(w | \alpha, \beta) = \int_{\theta} p(\theta | \alpha) \left( \prod_{n=1}^N \sum_{z_n} p(z_n | \theta) \right) d\theta. \quad (2.20)$$

$$\int_{\varphi} \prod_{j=1}^T p(\varphi_j | \beta) \left( \prod_{n=1}^N \sum_{z_n} p(w_n | \varphi_{z_n}) \right) d\varphi$$

Dokümanların marjinal olasılıkları çarpımını kullanarak yapının genel olasılığı eşitlik (2.21)'de elde edilir.

$$p(D | \alpha, \beta) = \prod_{d=1}^D \int_{\theta_d} p(\theta_d | \alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) \right) d\theta_d. \quad (2.21)$$

$$\int_{\varphi} \prod_{j=1}^T p(\varphi_j | \beta) \left( \prod_{d=1}^D \prod_{n=1}^{N_d} \sum_{z_{dn}} p(w_{dn} | \varphi_{z_{dn}}) \right) d\varphi$$

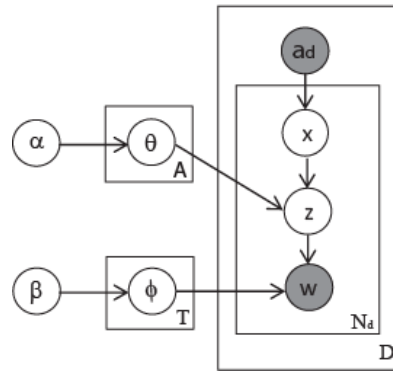
Şekil 2.6’da GDT üç aşamalı olasılıklı bir model olarak gösterilmiştir.  $\alpha$  ve  $\beta$  parametreleri yapı aşamasındaki parametrelerdir. Bu parametrelerin üretme sürecinde bir kez örneklendiği düşünülür.  $\theta_d$  değişkeni, doküman seviyesindeki değişkendir ve her doküman için bir kez örneklenir. Son olarak  $z_{dn}$  ve  $w_{dn}$  kelime seviyesindeki değişkenlerdir ve her doküman içindeki her kelime için bir kez örneklenir.

Sonlu bir rassal değişken kümesi olan  $\{z_1, \dots, z_N\}$  için eğer ortak dağılım sabit ise değişebilir denilir. GDT modelinde kelimelerin konulardan üretildiği varsayılmaktadır. Bu konular ise bir doküman içerisinde rassal bir sırada dağılır. Kelimelerin ve konuların sırasının olasılığı eşitlik (2.22)’de verilmektedir.

$$p(w, z) = \int p(\theta) \left( \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n) \right) d\theta \quad (2.22)$$

### 2.3.6. Yazar-Konu Modeli

Yazar-Konu (YK) modeli [1], GDT modelinin dokümanları yazan yazarları içine alan genişletilmiş bir sürümüdür. Her dokümanın yerine her dokümanın yazarı, bir konu dağılımına tahsis edilir ve her konu bir kelime dağılımına tahsis edilir. Bu durumda sadece dokümanların değil yazarların ilgi alanı da öğrenilebilir. Bir  $d$  dokümanı yazan yazar ya da yazar grubu  $a_d$  olarak ifade edildiği durumda, bir doküman içindeki her bir kelime için  $a_d$ ’den bir yazar seçilir. Bu yazara bir konu ve o konuya bir kelime dağılımı tahsis edilir.



Şekil 2.7: YK modelinin plaka gösterimi.

Modelin plaka gösterimi Şekil 2.7’de verilmektedir. Şekilde,  $x$  seçilmiş bir kelime için sorumlu bir yazarı temsil eder. Her yazar bir konu dağılımı  $\theta$  ile ilişkilendirilir. Bu model aşağıdaki süreci kullanır [1].

- Adım 1: Her bir yazar  $\alpha = 1, \dots, A$  için  $\theta$  seç,  $\theta_a \sim \text{Dir}(\alpha)$
- Adım 2: Her bir konu  $t = 1, \dots, T$  için  $\varphi$  seç,  $\varphi_t \sim \text{Dir}(\beta)$
- Adım 3: Her bir doküman  $d = 1, \dots, D$  için
- Adım 3.1:  $i = 1, \dots, N_d$  ile indekslenmiş her bir kelime  $w_i$  için
- Adım 3.1.1:  $a_d$ ’ye bağlı olarak bir yazar  $x_i$ ’yi seç
- Adım 3.1.2:  $x_i$ ’ye bağlı olarak bir konu  $z_i$ ’yi  $\theta_{x_i}$ ’den seç
- Adım 3.1.3:  $z_i$ ’ye bağlı olarak  $x_i$  bir kelime  $w_i$ ’yi  $\varphi_{z_i}$ ’den seç

$\theta$  ve  $\varphi$ , tahmin edilerek, yazarların yazdığı dokümanların içeriği ile ilgili ve dokümanlardan çıkarılan konu başlıkları ile ilgili bilgi edinilmesi hedeflenmiştir. Yapının genel olasılığı eşitlik (2.23)’te verilmektedir.

$$p(D|\theta, \varphi, A) = \prod_{d=1}^D p(w_d | \theta, \varphi, a_d) \quad (2.23)$$

Tek bir dokümanın olasılığı eşitlik (2.24)’de verilmektedir.

$$\begin{aligned} p(\mathbf{w}|\theta, \varphi, A) &= \prod_{n=1}^N p(w_n | \theta, \varphi, a_d) \\ &= \prod_{n=1}^N \sum_{\alpha=1}^A \sum_{j=1}^T p(w_n, z_n = j, x_n = \alpha | \theta, \varphi, a_d) \\ &= \prod_{n=1}^N \sum_{\alpha=1}^A \sum_{j=1}^T p(w_n | z_n = j, \varphi_j) p(z_n = j | x_n = \alpha, \theta_a) p(x_n = \alpha | a_d) \quad (2.24) \\ &= \prod_{n=1}^N \frac{1}{A_d} \sum_{a \in a_d} \sum_{j=1}^T p(w_n | z_n = j, \varphi_j) p(z_n = j | x_n = \alpha, \theta_a) \\ &= \prod_{n=1}^N \frac{1}{A_d} \sum_{x_n} \sum_{z_n} p(w_n | \varphi_{z_n}) p(z_n | \theta_{x_n}) \end{aligned}$$

Tek bir dokümanın olasılığını elde etmek için  $\theta$  ile  $\varphi$ 'nin ortak olasılık dağılımı ve integrali kullanılır. Eşitlik (2.25) ve eşitlik (2.26)'da verilmektedir.

$$p(\mathbf{w}|\alpha, \beta, A) = \iint \prod_{\alpha=1}^A p(\theta_a | \alpha) \cdot \prod_{j=1}^T p(\varphi_j | \beta) \cdot \prod_{n=1}^N \frac{1}{A_d} \sum_{x_n} \sum_{z_n} p(w_n | \varphi_{z_n}) p(z_n | \theta_{z_n}) d\theta d\varphi \quad (2.25)$$

$$\begin{aligned} p(D|\alpha, \beta, A) &= \int \prod_{\theta} \prod_{\alpha=1}^A p(\theta_a | \alpha) \prod_{d=1}^D \prod_{n=1}^{N_d} \sum_{x_{dn}} \sum_{z_{dn}} p(z_{dn} | \theta_{x_{dn}}) d\theta \\ &= \int \prod_{\varphi} \prod_{j=1}^T p(\varphi_j | \beta) \prod_{d=1}^D \prod_{n=1}^{N_d} \sum_{z_{dn}} \sum_{x_{dn}} p(w_{dn} | \varphi_{z_{dn}}) d\varphi \end{aligned} \quad (2.26)$$

Modelin değişebilirlik özelliğine göre bir tek doküman için  $w$ ,  $z$  ve  $x$ 'in ortak olasılık dağılımı eşitlik (2.27)'de verilmektedir.

$$p(\mathbf{w}, z, x|\alpha, \beta, A) = \iint \prod_{\alpha=1}^A p(\theta_a | \alpha) \cdot \prod_{j=1}^T p(\varphi_j | \beta) \cdot \prod_{n=1}^N \frac{1}{A_d} p(w_n | \varphi_{z_n}) p(z_n | \theta_{z_n}) d\theta d\varphi \quad (2.27)$$

### 2.3.7. Gibbs Örnekleme

Her iki model için en temel konu doküman-konu dağılımı  $\theta$  ve konu-kelime dağılımı  $\varphi$ 'yi elde etmektir. Hoffman dağılımı doğrudan elde etmek için beklenen maksimum algoritmasını kullanmıştır [33]. Fakat bu algoritmanın olabilirlik fonksiyonun yerel maksimum problemine sahip olması nedeniyle, Gibbs Örnekleme yöntemi kullanılmaya başlanmıştır. Gibbs örnekleme, GDT modelinin dağılımını tahmin etmek için kullanılan basit ve faydalı bir yöntemidir [35]. Farklı konular içeren bir doküman, her bir konuyu anlatmak için kullanılan bir kelime dağılımı olsun.  $T$  sayıda konu olduğunda, bir kelimenin olasılığı eşitlik (2.28)'de verilmektedir.

$$p(w_i) = \sum_{j=1}^T p(w_i | z_i = j) p(z_i = j) \quad (2.28)$$

İlgili oldukları konu için önemli ise  $p(w|z)$  değeri yüksektir. Burada  $w_i$  değeri kelime indekslerinin kümesidir.  $p(z)$  ise bir doküman içerisindeki konular arasında en yaygın olanı yani egemen olanıdır. Bu yöntemde hem kelimeler hem de konular için bir sözlük  $V$  kullanılmaktadır. Dokümanlar içerisinde geçen her kelime bu sözlük içerisinde vardır. Bu yöntem direk kelimelerle değil, kelimelerin sözlükte geçen indeksleri ile ilgilenir.  $V$  sözlüğü üzerinden  $p(w|z=j)$ ,  $T$  katlı dağılımın bir kümesidir ve  $p(w|z=j) = \varphi_w^j$ 'dir.  $p(z)$  ise  $T$  konu üzerinden  $D$  katlı dağılımın bir kümesidir. Bir  $d$  dokümanı içerisindeki bir kelime için konu olasılıkları  $p(z=j) = \theta_j^d$ 'dir.  $\theta$  ve  $\varphi$  değerlerini elde etmek için sonsal dağılımı analiz etmek gerekir. Gibbs örnekleme sıra ile bütün şartlı dağılımlardan örnekleme yaparak modeldeki parametrelerin ortak sonsal yoğunluk dağılımına örnekler verecektir. Bu yöntem  $p(z_i = j | z_{\setminus i}, w_i, d_i, .)$  olasılığının hesaplamasına bağlıdır. Eğer her doküman için  $z$  değeri bilinirse  $\theta$  ve  $\varphi$ 'yi hesaplamak basitleşecektir. Burada  $z_i = j$ ,  $j$  konusuna  $i$  kelimesinin tahsisi ve  $z_{\setminus i}$  diğer tüm kelimelerin konu tahsisidir. “.” ise diğer parametreleri temsil etmektedir. Griffiths and Stevvers ait olasılık hesabı eşitlik (2.29)'da verilmektedir.

$$p(z_i = j | z_{\setminus i}, w_i, d_i, .) \propto p(w_i) = \sum_{j=1}^T p(w_i | z_i = j) p(z_i = j) \quad (2.29)$$

$C^{WT}$  ve  $C^{DT}$  sırasıyla  $W \times T$  ve  $D \times T$  boyutları ile sayıların matrisidir.  $C_{wj}^{WT}$  o anki örnek olan  $i$ 'yi içermeyen konu  $j$ 'ye kelime  $w$ 'nin kaç kez tahsis edildiği sayısını verir.  $C_{dj}^{DT}$  o anki örnek olan  $i$ 'yi içermeyen doküman  $d$  içerisindeki bazı kelimelere konu  $j$ 'nin kaç kez tahsis edildiği bilgisini verir. Denklemdaki sağ bölüm doküman  $d$  için o anki konu dağılımı altında konu  $j$ 'nin olasılığını verir. Sol bölüm ise konu  $j$  altındaki kelime  $w$ 'nin olasılığını verir. Belgelerin genelinde, algoritmanın çalışması sırasında bir kelimenin bir konuya birden fazla kez tahsisi gerçekleşmişse, algoritmanın bitiminde bu kelimenin bu konuya tahsis olasılığı

artacaktır. Aynı zamanda, eğer konu  $j$  bir doküman içerisinde birçok kez kullanılmış ise, bu doküman konu  $j$  ile ilişkilendirileceğinden doküman içerisinde geçen her kelimenin bu konuya tahsis edilme olasılığı artacaktır.

Gibbs örnekleme algoritması  $[1...T]$  aralığından rastgele bir konu seçer ve bu kelimelerin her birini bu konuya tahsis etmeye başlar. İlk olarak, her kelime için  $C^{WT}$  ve  $C^{DT}$  matrisleri o anki konu tahsisinden dolayı bir azalır. Sonra dağılımdan yeni bir konu örneklenir ve matrisler bu yeni konu tahsisi ile artırılırlar. Örnekleme sürecinin başlangıç seviyesi sırasında, Gibbs örnekleri sonsal dağılıma ait zayıf tahminlerde bulunduğu için atılırlar. Bu sürece yanma periyodu da denir. Bu periyoddan sonra, başarılı Gibbs örnekleri hedef dağılıma yaklaştırmaya başlar. Bu dağılımı temsil edecek örnekler bu aşamada belli aralıklarla kaydedilirler. Algoritma nihayetinde her kelime için  $z$  'ye ait hesaplamaları verecektir.  $z$  değeri kullanılarak, bu örneklerin her biri ile  $\theta$  ve  $\varphi$  değerleri eşitlik (2.30)'da olduğu gibi elde edilmektedir.

$$\varphi_i^j = \frac{C_{ij}^{WT} + \beta}{\sum_{l=1}^W C_{lj}^{WT} + V\beta} \quad \theta_j^d = \frac{C_{dj}^{DT} + \alpha}{\sum_{l=1}^T C_{dl}^{DT} + T\alpha} \quad (2.30)$$

Yazar-konu modeli de GDT modeli gibi Gibbs örnekleme yöntemini kullanır [34]. YK modele ait denklem eşitlik (2.31)'de verilmektedir.

$$p(z_i = j, x_i = k \mid w_i = m, z_{\setminus i}, x_{\setminus i}, w_{\setminus i}, a_d) \propto \frac{C_{mj}^{WT} + \beta}{\sum_{m'} C_{m'j}^{WT} + V\beta} \frac{C_{kj}^{AT} + \alpha}{\sum_{j'} C_{kj'}^{AT} + T\alpha} \quad (2.31)$$

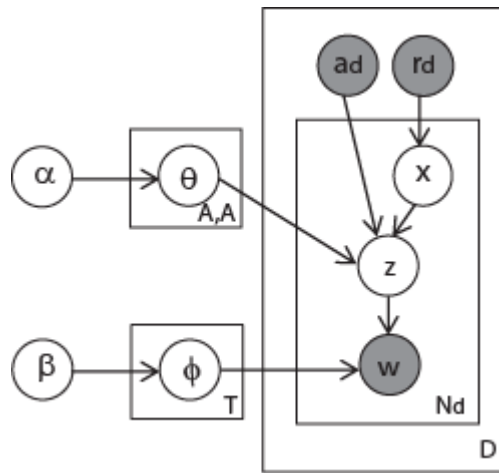
Bu denklemde  $z_i = j$  konu  $j$  'nin bir doküman içerisindeki  $i$  . kelimeye tahsisi anlamına gelmektedir.  $x_i = k$  eşitliği ise yazar  $k$  'nın aynı  $i$  . kelimeye tahsisidir.  $w_i = m$  ,  $i$  . kelimenin kullanılan sözlükte,  $m$  . kelime olduğu anlamına gelir.  $z_{\setminus i}$  ve  $x_{\setminus i}$  o anki değerlendirilen konu ve yazar hariç bütün konu ve yazarların tahsisi anlamına gelmektedir.  $C_{mj}^{WT}$  kelime  $m$  'nin konu  $j$  'ye kaç kez tahsis edildiğinin sayısını,  $C_{kj}^{AT}$  ise yazar  $k$  'nın konu  $j$  'ye kaç kez atandığının sayısını verir. Eşitlik (2.32) ile Yazar-Konu modelinde bir konuya tahsis edilen kelimelerin olasılığı ve o

konuya tahsis edilen yazarların olasılığı bulunur. Algoritma rastgele olarak kelimeleri konu ve yazarlara tahsis ederek çalışmasına başlar ve her iterasyonda daha başarılı örnekler elde eder ve kaydetmeye başlar. Bu iterasyonların sayısı önceden bellidir ve sisteme girilir. Yanma periyodundan sonra, örnekler alınır ve  $\theta$  ve  $\varphi$  hesaplanır.

$$\varphi'_{mj} = \frac{C_{mj}^{WT} + \beta}{\sum_{m'} C_{m'j}^{WT} + V\beta} \quad \theta^{\text{vd}}_j = \frac{C_{kj}^{AT} + \alpha}{\sum_{j'} C_{kj'}^{AT} + T\alpha} \quad (2.32)$$

### 2.3.8. Yazar-Alıcı-Konu Modeli

Yazar-Alıcı-Konu (YAK) modeli [13] bir sosyal ağ yapısına sahip mesajlardan oluşan veri kümesi için geliştirilmiş her bir mesajın göndericileri ve alıcıları arasındaki konu benzerliği olasılığını hesaplayan yönlü bir grafiksel modeldir. Yazar-Konu modeli baz alınarak geliştirilmiştir, fakat farklı olarak Yazar-Alıcı-Konu modelinde konuların keşfedilmesi işlemi sosyal ağ yapısından etkilenmektedir. Her konu-kelime dağılımı her yazar-alıcı çiftine ait konuların dağılımı üzerinden tanımlanır. Kişiler arasındaki yakınlığı ölçmede kullanılan etkin bir yöntem olarak kullanılmaktadır. Şekil 2.8’de modelin üretici süreci verilmektedir.



Şekil 2.8: YAK modelinin plaka gösterimi.

YAK modelinde her mesaj  $d$  için, bir yazar  $a_d$  ve bir alıcı kümesi  $r_d$  gözlenmektedir. Her bir kelimeyi üretmek için bir alıcı  $x$ ,  $r_d$  kümesinden ve sonra

bir konu  $z$  bir  $\theta_{a_d x}$  konu dağılımından seçilir.  $\theta_{a_d x}$  işlemde olan mesaja ait yazar-alıcı çiftine  $(a_d, x)$  özel bir dağılımdır. En son olarak, kelime  $w$   $\varphi_z$  dağılımdan üretilir.

Modelde Gibbs örnekleme sürecinin tamamlanması için bütün diğer kelimelerin  $x_{-di}$  alıcı ve  $z_{-di}$  konu tahsisleri verildiğinde  $d$  dokümanındaki  $i$ . kelime olan  $w_{di}$  için bir konunun ve alıcının koşullu dağılımının  $p(x_{di}, z_{di} = k | x_{-di}, z_{-di}, w, \alpha, \beta, a, r)$  hesaplanması eşitlik (2.33)'te verilmektedir.

$$p(x_{di}, z_{di} = k | x_{-di}, z_{-di}, w, \alpha, \beta, a, r) \quad (2.33)$$

$$\propto \frac{\alpha_{z_{di}} + n_{a_d x_{di} z_{di}} - 1}{\sum_{t=1}^T (\alpha_t + n_{a_d x_d t}) - 1} \frac{\beta_{w_{di}} + m_{z_{di} w_{di}} - 1}{\sum_{v=1}^V (\beta_t + m_{z_{di} v}) - 1}$$

$n_{ijt}$ ,  $t$  konusunun  $(i, j)$  alıcı-yazar çiftine tahsis edilme sayısı ve  $m_{tv}$ ,  $v$  kelimesinin  $t$  konusuna tahsis edilme sayıdır.  $\theta$  ve  $\varphi$ 'nin verilen bir veri kümesindeki sonsal hesabı eşitlik (2.34) ve eşitlik (2.35)'de verilmektedir.

$$\theta_{ijz} = \frac{\alpha_z + n_{ijz}}{\sum_{v=1}^V (\alpha_t + n_{ijt})} \quad (2.34)$$

$$\theta_{ijz} = \frac{\alpha_z + n_{ijz}}{\sum_{v=1}^V (\alpha_t + n_{ijt})} \quad (2.35)$$

## 2.4. Topluluk Tanıma Modelleri

Son zamanlarda, sosyal ağ sitelerinin popülaritesi hızlı bir şekilde artmıştır. İnsanlar sosyal ağların sağladığı birçok servisten yararlanmakta, yeni insanlarla tanışmakta, eski arkadaşları ile iletişime geçmekte, düşüncelerini ve ilgi alanlarını paylaşmaktadırlar. Ticaret, akademi ve güvenlik gibi birçok alanda merak uyandıran insanlar arasındaki tartışılan konuların belirlenmesi, ağ üyeleri arasındaki ilişkilerin

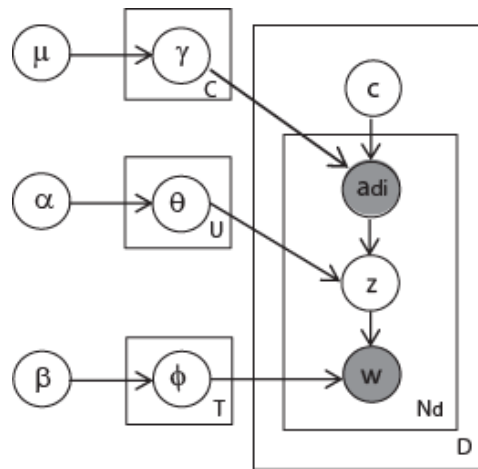
anlaşılması problemleri sosyal ağ analiz konusunun gelişmesine katkıda bulunmaktadır. Sosyal ağ analizinin en önemli problemi topluluk tanıma konusudur.

Bir sosyal ağda belirli ölçütlere göre birbirine yakın olan insanların oluşturduğu öbeğe topluluk denir. Benzer bağlantılara, paylaşılan ortak konulara, yakın konumlara göre insanlar topluluklara ayrılabilirler. Sosyal ağ birden fazla topluluğa ayırtmak ağın analizini kolaylaştırmaktadır. Aynı topluluğa ait olan insanların ortak ilgi alanlarına sahip olduğu ön görüldüğünde kişilerin ve kişiler arasındaki bağlantıların analizi de kolaylaşmaktadır. Bu bölümde şüpheli toplulukların analizi için topluluk tanıma yöntemleri incelenmiştir.

### 2.4.1. Topluluk-Kullanıcı-Konu Modeli

Topluluk-Kullanıcı-Konu (TKK) modeli Yazar-Konu modeli baz alınarak geliştirilmiştir. Sosyal ağların topluluk yapısı, bu modelde sosyal ağdaki üyelerinin arasındaki iletişim dokümanları kullanılarak analiz edilmektedir. Her bir  $u$  kullanıcısı için  $c$  topluluğuna ait olma şartlı olasılığı  $P(u|c)$  tanımlanmıştır. Daha sonra, kullanıcılar kelimelerin dağılımından oluşan bir konu kümesi ile etiketlenmektedir.

Şekil 2.9'da TKK modeli için Bayes ağının hiyerarşi gösterilmektedir. Yazar-Konu modelinde olduğu gibi  $\theta$  kullanıcı-konu dağılımını ve  $\varphi$  konu-kelime dağılımını göstermektedir. Bu modelde, her bir  $c$  topluluğu için kullanıcılardan oluşan  $\gamma$  dağılımı tanımlanmıştır.



Şekil 2.9: TKK modelinin plaka gösterimi.

Model içindeki bütün değişkenlerin birleşik dağılımının sonucu eşitlik (2.36), (2.37) ve (2.38)'de gösterilmektedir.  $P(c, u, z | w)$ , belli bir  $c$  topluluğuna,  $u$  kullanıcısına ve  $z$  konusuna her bir  $w$  kelimesinin tahsis edilmesinin sonsal dağılımıdır. TKK modeli için Gibbs Örnekleme,  $w_i, z_{-i}, x_{-i}$ 'nin şu anki durumu hariç bütün kelime tahsislerine bağlı olarak  $r$  konusuna ait  $p$  topluluğu ve  $q$  kullanıcısı tarafından üretilen  $w_i$ 'nin olasılığı  $P(c_i = p, u_i = q, z_i = r | w_i = m, z_{-i}, x_{-i}, w_{-i})$  hesaplanması ile gerçekleşir. TKK modeli için  $P(c_i = p, u_i = q, z_i = r | w_i = m, z_{-i}, x_{-i}, w_{-i})$  olasılığı hesabı eşitlik (2.36)'da verilmektedir.

$$P(w_i = m, z_i = r) \propto \frac{C_{mr}^{WT} + \beta}{\sum_{m'} C_{m'r}^{WT} + V\beta} \quad (2.36)$$

$$P(z_i = r | u_i = q) \propto \frac{C_{rq}^{TU} + \alpha}{\sum_{r'} C_{r'q}^{TU} + T\alpha} \quad (2.37)$$

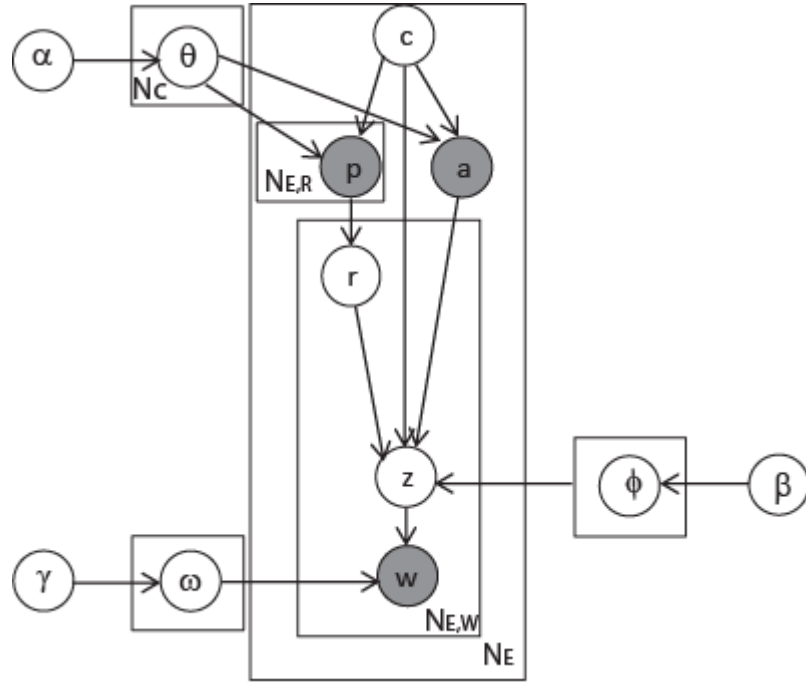
$$P(u_i = q | c_i = p) \propto \frac{C_{qp}^{UC} + \gamma}{\sum_{q'} C_{q'p}^{UC} + U\gamma} \quad (2.38)$$

Yukardaki denklemden,  $C_{mr}^{WT}$   $w_i = m$  kelimesinin  $z_i = r$  konusuna,  $C_{rq}^{TU}$   $z_i = r$  konusunun  $u_i = q$  kullanıcısına ve  $C_{qp}^{UC}$  ise  $u_i = q$  kullanıcısının  $c_i = p$  topluluğuna tahsis edilme sayısını vermektedir.

## 2.4.2. Topluluk-Yazar-Alıcı-Konu Modeli

Topluluk-Yazar-Alıcı-Konu (TYAK) modeli, sosyal ağlarda konu ve bağlantı bilgisini kullanan bir olasılıklı modeldir. TKK ve YAK model temel alınarak sosyal ağlardaki toplulukları tanımlamaya çalışmaktadır. Sosyal ağlarda mesajları gönderen kişi ve alıcı bilgisi YAK modelde olduğu gibi kullanılmıştır. Doküman kümesi ağ kullanıcıları arasındaki e-postalardan oluşmaktadır. Kullanıcılara ait e-posta kümesi

verildiğinde, Topluluk-Yazar-Alıcı-Konu modeli, yapıdaki konuşulan sosyal konuları tespit ettiği gibi ağdaki gizil topluluklarında keşfedilmesine yardımcı olur.



Şekil 2.10: TYAK modelinin plaka gösterimi.

Şekil 2.10’da modelin yapısı verilmektedir. Her  $e_d$  e-postası için, eşit oranda dağılacak şekilde rastgele bir  $c_d$  topluluğu,  $c_d$  topluluğu için o e-postaya ait  $a_d$  yazarlar ve  $p_d$  alıcılar kümesi seçilmektedir. E-postadaki her bir kelimeyi üretmek için  $p_d$  alıcılar kümesinden eşit oranda dağılacak şekilde rastgele bir  $r$  alıcısı seçilmektedir.  $c_d$ ,  $a_d$  ve  $r$  için bir  $z$  konusu seçilir ve  $z$  konusuna bağlı olarak  $w_{di}$  kelimesi üretilir.

Bir  $a_d$  e-postası için birleşik olasılık dağılımı eşitlik (2.39)’da verilmektedir.

$$P(c_d, a_d, p_d, r_d, z_d, w_d) = p(c_d) p(a_d | c_d) \prod_{r \in p_d} p(r | c_d) \prod_{i=1}^{N_d} p(w_{di} | z_{di}) p(z_{di} | c_d, a_d, r_{di}) \quad (2.39)$$

$r$ ,  $p_d$  alıcı kümesinden seçilen gizil alıcı,  $z_d$   $d$  e-postası içerisindeki  $w_d$  kelime dizisine karşılık gelen gizil konu dizisidir.  $z_{di}$ ,  $i$ . kelimeye karşılık düşen gizil konudur.

Modelin topluluk tahsisinin koşullu dağılımı Gibbs Örnekleme hesabı eşitlik (2.40)'da olduğu gibidir.

$$p(c_d = c | c_{-d}, p, r, a, z, w) \propto \prod_{u_i \in \{a_d, p_d\}} (n_{-d, cu_i}^{CU} + \alpha) \times \prod_{r \in p_d} \left( \frac{\prod_{z=1}^T \Gamma(e_{d, rz} + n_{-d, (c_d a_d r)z}^{(CUU)T}) + \beta}{\Gamma(\sum_{z=1}^T e_{d, rz} + n_{-d, (c_d a_d r)z}^{(CUU)T}) + T\beta} \right) \prod_{i=0}^{|p_d|} \sum_{u=1}^U (n_{-d, cu}^{CU} + U\alpha + i) \quad (2.40)$$

$n_{-d, cu_i}^{CU}$  kullanıcı  $u_i$ 'nin  $d$  e-postası haricinde kaç kez  $c$  topluluğuna atandığını gösterir.  $e_{d, rz}$  konu  $z$ 'nin  $d$  e-postası içinde  $r$  alıcısına ve  $n_{-d, (c_d a_d r)z}^{(CUU)T}$  konu  $z$ 'nin  $d$  e-postası haricinde  $(c_d a_d, r)$  topluluğu, yazarı ve alıcısına kaç kez tahsis edildiğinin sayısını verir. Bir  $w_{di}$  kelimesi için alıcı-konu ögesinin şartlı dağılımı eşitlik (2.41)'de verilmektedir.

$$p(r_{di} = r, z_{di} = z | c_{-d}, a_{-d}, p_{-d}, r_{-di}, z_{-di}, w_{-di}, c_d, a_d, p_d, w_{di} = w) \propto \frac{n_{-di, zw}^{TW} + \gamma}{\sum_{v=1}^W n_{-di, zv}^{TW} + W\gamma} \times \frac{n_{-di, (c_d a_d r)z}^{(CUU)T} + \beta}{\sum_{h=1}^T n_{-di, (c_d a_d r)h}^{(CUU)T} + T\beta} \quad (2.41)$$

$n_{-di, xy}^{XY}$ ,  $d$  e-postası içinde  $i$ . örneğin dışında  $x \in X$  tarafından  $y \in Y$ 'nin kaç kez üretildiğinin sayısını verir.

Bu modelde bir kullanıcı birden fazla topluluğa tahsis edilebilir. Topluluklara tahsis edilme sayılarına göre bireyler arasındaki en popüler kullanıcılar ve bireyler arasındaki en popüler konular bulunabilir.

### 2.4.3. Topluluk-Konu Modeli

Topluluk-Konu (TK) modeli ilgi alanlarındaki benzerliğin yüksek olduğu kişileri aynı topluluk altında kümelemeye çalışır. Topluluğun konu dağılımındaki benzerlik temel alınır. TK modeli, bütün yazarların konu dağılımlarını baz alarak toplulukları oluşturur ve benzer konu dağılımlarına sahip olan yazarları bir topluluk

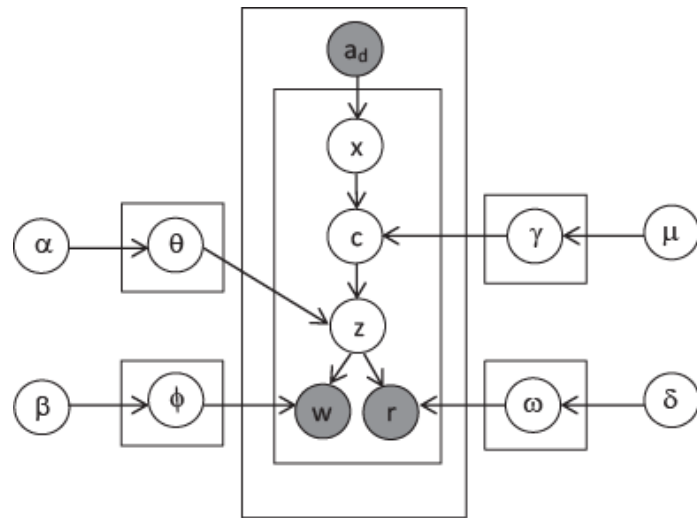
altında kümeler. Diğer topluluk tanıma yöntemlerinde süreç içerisinde topluluk tahsisi doküman seviyesinde iken TK modeli topluluk tahsisini Şekil 2.11’de görüldüğü gibi konu seviyesinde yapmaktadır. Bu durum benzer ilgi alanlarına ve ortak konulara sahip olan kişilerin çok daha doğru bir şekilde aynı topluluk altına kümelenmesini sağlar.

Şekil 2.11’de modelin öğrenme sürecinin grafiksel temsili bulunmaktadır.  $d$  dokümanındaki her bir  $x$  yazarı için,  $x$  öncelikle yazar-topluluk dağılımı  $\gamma$ ’dan bir  $c$  topluluğu seçer; sonra konu-topluluk dağılımı  $\theta$ ’dan  $c$  topluluğu altında bir  $z$  konusu seçer. Son olarak konu-kelime dağılımı  $\phi$ ’dan  $z$  konusu altında bir  $w$  kelimesi seçer. Bu modeldeki  $a_d$ ,  $w$  ve  $r$  gözlenen değişkenlerdir. Diğer değişkenler ise gizlidir.

Modelin Gibbs örnekleme sürecinde 3 parametre tahmini yapılır. Bunlar:  $\phi$ ,  $\theta$  ve  $\gamma$ ’dir. Bu parametrelerin başlangıç değerleri sırasıyla  $\beta = 0.01$ ,  $\alpha = 50/T$  ve  $\mu = 50/C$  olarak verilir.

1000 tekrarlama sonra  $\phi$ ,  $\theta$  ve  $\gamma$ ’nin son değerleri elde edilir. Her tekrarlama da,  $d$  dokümanı içerisindeki her  $w$  kelimesi ve  $d$  dokümanına ait her yazar için bir  $z$  konusu ve  $c$  topluluğu tahsis edilir. Her olası  $(c, z) \in C \times T$  ataması için, olasılık eşitlik (2.42)’de olduğu gibi hesaplanır.

$$P_{c,z} \left( z^i = z, c^i = c, x^i = x | w^i = w, z^{-i}, x^{-i}, y^{-i}, w^{-i}, a_d \right) \propto \gamma^{i-1} \cdot \phi^{i-1} \cdot \theta^{i-1} \quad (2.42)$$



Şekil 2.11: TK modelinin plaka gösterimi.

$i$  ve  $i-1$  o anki ve ondan önceki tekrarlamayı temsil eder.  $i$ . tekrarlama da bütün atamalardan sonra, olasılık matrisleri eşitlik (2.43), (2.44) ve (2.45)'te olduğu gibi güncellenir.

$$\gamma_{x,c} = \frac{n_{x,c} + \mu}{\sum_{c' \in \cdot} (n_{x,c'} + \mu)} \quad (2.43)$$

$$\theta_{c,z} = \frac{n_{c,z} + \alpha}{\sum_{z' \in T} (n_{c,z'} + \alpha)} \quad (2.44)$$

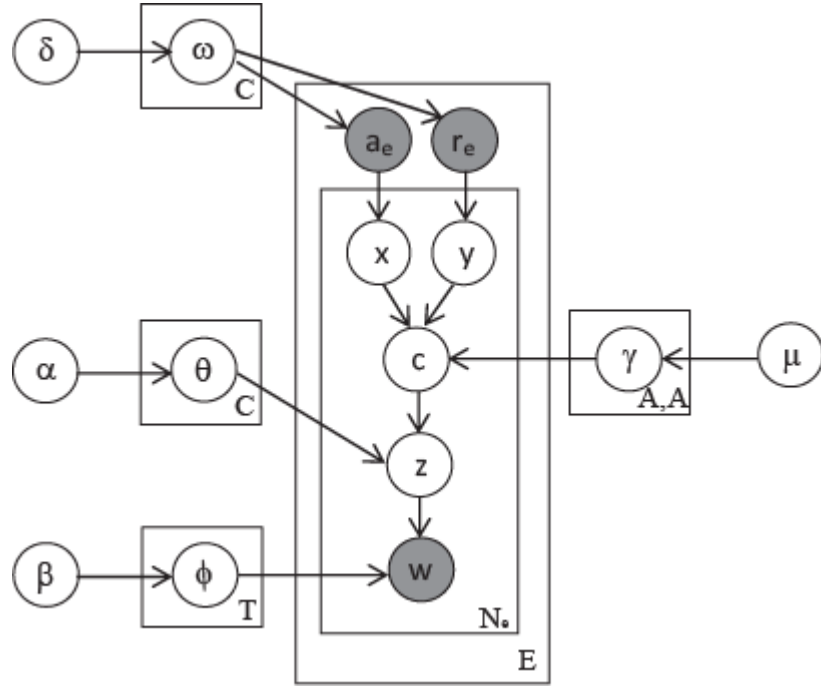
$$\phi_{z,w} = \frac{n_{z,w} + \beta}{\sum_{w' \in V} (n_{z,w'} + \beta)} \quad (2.45)$$

Yukardaki denklemlerde  $n_{x,c}$  o anki tekrarlama  $x$  yazarının  $c$  topluluğuna kaç kez tahsis edildiği bilgisini tutar.  $\gamma_{x,c}$  ise  $\gamma$  matrisindeki  $x$  yazarına ve  $c$  topluluğuna ait olasılık bilgisini tutar. Benzer ifade  $\theta_{c,z}$  ve  $\phi_{z,w}$  için de yazılabilir.

### 3. GELİŞTİRİLEN MODEL

#### 3.1. Yazar-Alıcı-Topluluk-Konu Modeli

Yazar-Alıcı-Topluluk-Konu (YATK) modeli kullanıcıların birbirlerine göndermiş olduğu mesaj içeriğinin benzerliğine göre yazar-alıcı çiftlerini topluluklara tahsis eder. Benzer konu dağılımlarına sahip yazar-alıcı çiftleri aynı topluluk altında kümelenir. YATK modeli, TK modeli temel alınarak oluşturulmuştur. YATK modeli, sosyal ağdaki düğüm bilgisinin yanında düğümler arasındaki kenar bilgisini de toplulukları oluşturmak için kullanmaktadır.



Şekil 3.1: YATK modelinin plaka gösterimi.

Şekil 3.1’de modelin öğrenme sürecinin grafiksel temsili bulunmaktadır. YATK modelindeki dokümanlar e-posta olarak tanımlanmıştır. Modelde her bir  $e$  e-postası, o e-postaya ait  $a_e$  yazarlar kümesi,  $r_e$  alıcılar kümesi ve e-postadaki  $w_e$  kelime kümesi gözlenen değişkenlerdir.  $e$  dokümanındaki her bir  $x$  ve rastgele seçilmiş bir  $y$  yazarı için öncelikle yazar-alıcı-topluluk dağılımı  $\gamma$ ’dan bir  $c$  topluluğu seçilir; sonra konu-topluluk dağılımı  $\theta$ ’dan  $c$  topluluğu altında bir  $z$

konusu seçilir. Son olarak konu-kelime dağılımı  $\varphi$ ’dan  $z$  konusu altında bir  $w$  kelimesi seçilir.

Gibbs örnekleme sürecinde 3 parametre tahmini yapılır. Bunlar:  $\varphi$ ,  $\theta$  ve  $\gamma$ ’dir. Bu parametrelerin başlangıç değerleri sırasıyla  $\beta = 0.01$ ,  $\alpha = 50/T$  ve  $\mu = 50/C$  olarak verilir. 1000 tekrarlama sonra  $\varphi$ ,  $\theta$  ve  $\gamma$ ’nin son değerleri elde edilir. Her tekrarlama, YATK modeli  $e$  e-postası içerisindeki her  $w$  kelimesi ve  $e$  dokümanına ait her yazar-alıcı çifti için bir  $z$  konusu ve  $c$  topluluğu tahsis eder. YATK modelinin Gibbs örnekleme sürecindeki ana adımı,  $P_{c,z}(z^i, c^i, x^i, y^i | w^i)$  olasılığının tahminidir. Her tekrarlama olasılık eşitlik (3.1)’de olduğu gibi hesaplanır

$$P_{c,z}(z^i = z, c^i = c, x^i = x, y^i = y | w^i = w, z^{-i}, x^{-i}, y^{-i}, w^{-i}, a_e, r_e) \propto \gamma^{i-1} \cdot \varphi^{i-1} \cdot \theta^{i-1} \quad (3.1)$$

Denklemden  $i$  ve  $i-1$  o anki ve ondan önceki tekrarlamaı temsil eder.  $i$ . tekrarlama bütun atamalardan sonra, olasılık matrisleri eşitlik (3.2), eşitlik (3.3) ve eşitlik (3.4)’de olduğu gibi güncellenir.

$$\gamma_{xy,c} = \frac{n_{xy,c} + \mu}{\sum_{c' \in C} (n_{xy,c'} + \mu)} \quad (3.2)$$

$$\theta_{c,z} = \frac{n_{c,z} + \alpha}{\sum_{z' \in T} (n_{c,z'} + \alpha)} \quad (3.3)$$

$$\varphi_{z,w} = \frac{n_{z,w} + \beta}{\sum_{w' \in V} (n_{z,w'} + \beta)} \quad (3.4)$$

Yukardaki denklemden  $n_{xy,c}$  o anki tekrarlama x yazarının ve y alıcısının  $c$  topluluğuna kaç kez tahsis edildiği bilgisini tutar.  $\gamma_{xy,c}$  ise  $\gamma$  matrisindeki  $xy$  yazar-alıcı çifti ve  $c$  topluluğu için olasılık bilgisini tutar. Benzer ifade  $\theta_{c,z}$  ve  $\varphi_{z,w}$  için de yazılabilir.

Kullanıcı-topluluk dağılım olasılıkları ise eşitlik (3.5)'te olduğu gibi hesaplanır.

$$p(u|c) = \frac{n_{cu} + \omega}{\sum_{c' \in C} (n_{cu'} + \omega)} \quad (3.5)$$

$n_{cu}$ ,  $u$  kullanıcısının  $c$  topluluğuna atanma sayısıdır ve  $p(u|c)$  o kullanıcının  $c$  topluluğa ait olma olasılığını vermektedir.

Aşağıda modelin algoritması verilmektedir.

- Adım 1: Her  $e$  e-postası için
  - Adım 1.1.: Her  $e$  e-postasının  $x$  yazarı için
    - Adım 1.1.1:  $e$  e-postasındaki her  $w_i$  kelimesi için
      - Adım 1.1.2:  $r_e$  alıcı kümesinden rastgele bir  $y$  alıcısı seç.
      - Adım 1.1.3:  $w_i$ 'ye,  $x$ 'e ve  $y$ 'ye rastgele bir topluluk ve konu ata.
- Adım 2:  $i = 1:1000$  sayısına
  - Adım 2.1: Her  $e$  e-postası için
    - Adım 2.1.1: Her  $e$  e-postasının  $x$  yazarı için
      - Adım 2.1.1.1:  $e$  e-postasındaki her  $w_i$  kelimesi için
        - Adım 2.1.1.2:  $r_e$  alıcı kümesinden rastgele bir  $y$  alıcısı seç.
        - Adım 2.1.1.3:  $P_{cz}$ 'yi hesapla ve  $w_i$ 'ye,  $x$ 'e ve  $y$ 'ye bir topluluk ve konu ata.
- Adım 3:  $\theta$ ,  $\gamma$  ve  $\phi$  matrislerini güncelle.

YATK model sosyal ağlardaki yazar-alıcı çifti arasındaki konuşulan konulara göre topluluklar oluşturur ve benzer konu ile ilgili konuşan tüm çiftleri bir topluluk altında sıralar. Bunun yanı sıra ağın bağlantı yapısının da topluluklar altında gözlenmesini sağlar. Model diğer modellerden farklı olarak etkili bir şekilde bağlantı analizi yapılabilmesini sağlamaktadır.

### 3.2. Analiz Yöntemleri

Bu çalışmada, toplulukları oluşturmak için konu ve bağlantı bilgisi kullanılmıştır. Bu nedenden dolayı, hem konu seviyesinde hem de bağlantı

seviyesinde karşılaştırma yapılmalıdır. Konu seviyesinde karşılaştırma yapmak için Kullback Leibler yöntemi, bağlantı seviyesinde karşılaştırma yapmak için ise normalleştirilmiş ortak bilgi ve modülerlik metodu kullanılmaktadır.

### 3.2.1. Kullback Leibler Yöntemi

KL, iki olasılık dağılımı arasındaki farkı veya benzerliği ölçmek için kullanılan oldukça yaygın bir yöntemdir.  $x_1$  ve  $x_2$  değişkenlerinin arasındaki benzerlik, bu değişkenlere ait  $\theta_{x_1}$  ve  $\theta_{x_2}$  konu dağılımları arasındaki benzerliğin ölçülmesi ile bulunmaktadır [36]. Yöntemin KL değeri eşitlik (3.6)'da verildiği gibi topluluk-konu dağılımları üzerinden hesaplanır. Bu çalışmada YATK modeli önceden belirlenen sayıda  $C = \{c_1, c_2, \dots\}$  topluluk oluşturmaktadır.  $c_1$  ve  $c_2$  toplulukları arasındaki benzerlik  $T$  sayıda konu verildiğinde aşağıdaki gibi hesaplanır.

$$KL(c_1, c_2) = \sum_{i=1}^T \left[ cz_{1i} \times \log \frac{cz_{1i}}{cz_{2i}} + cz_{2i} \times \log \frac{cz_{2i}}{cz_{1i}} \right] \quad (3.6)$$

KL değerinin yüksek olması topluluklar arasındaki konu benzerliğinin düşük seviyede olduğunu gösterir.

### 3.2.2. Normalleştirilmiş Ortak Bilgi Yöntemi

Bu çalışmada, Normalleştirilmiş Ortak Bilgi (NOB) yöntemi [37] bir topluluk içerisindeki yazar-alıcı çiftlerine ait kenarların gerçek sosyal ağdaki kenarlara ne kadar yakın olduğunu ölçmek için kullanılmaktadır.  $x = \{x_1, \dots, x_k\}$ ,  $k$  adet topluluğa ait kenar kümelerini ifade eder.  $x_k$ ,  $k$ . topluluğa ait kenar kümesini içerir. Asıl ağın tüm kenarlarına ait küme  $y$  olarak verildiğinde, ortak bilgi eşitlik (3.7)'de olduğu gibi hesaplanmaktadır.

$$OB(x, y) = \sum_{x_i, y} p(x, y) \log \frac{p(x, y)}{p(x) p(y)} \quad (3.7)$$

Eşitlik (3.6)'da elde edilen ortak bilgi değeri eşitlik (3.8)'de normalleştirilmiştir.

$$NOB(x, y) = \frac{OB(x, y)}{\max(H(x), H(y))} \quad (3.8)$$

$H(x)$  ve  $H(y)$ ,  $x$  ve  $y$  kümelerine ait entropi değerleridir. NOB yöntemi ile bulunan değer büyüdükçe model gerçek doğruya yaklaşmaktadır.

### 3.2.3. Modülerlik Yöntemi

Modülerlik, bir ağın ne kadar iyi parçalandığına dair bilgi veren bir kalite ölçütüdür [38]. Modülerlik değeri eşitlik (3.9)'da verilmektedir.

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - \|e^2\| \quad (3.9)$$

$Tre$  değeri  $j$  topluluğu içindeki düğümleri bağlayan ağa ait kenarların fraksiyonunu verir.  $\|x\|$ ,  $x$  matrisinin elemanlarının toplamını gösterir ve  $e$  matrisi bir  $k \times k$  simetrik matrisidir. Bu matrisin elemanı olan  $e_{ij}$  de  $i$  topluluğu içerisindeki düğümleri bağlayan ağın içerisindeki bütün kenarların fraksiyonudur.  $a_i = \sum_j e_{ij}$  ise topluluk  $i$  içerisindeki düğümleri bağlayan kenarların fraksiyonudur.

Hesaplama yapıldığında elde edilen yüksek modülerlik değeri ağın yüksek kalitede bölümlendiğini gösterir.

## 4. DENEYSEL ÇALIŞMALAR

### 4.1. Veri Kümesi

YATK modelin deney sonuçları Enron e-posta yapısı [39] üzerinde test edilmiştir. Enron Aralık 2001’de iflasını isteyen ABD’nin en büyük enerji ve dağıtım şirketi idi. Büyük şaşkınlığa yol açan iflas isteminden kısa bir zaman sonra şirketin yıllık faaliyet raporlarında büyük muhasebe hileleri yaparak halkını aldatılmış olduğu anlaşıldı. Bu yapı şüpheli sosyal bir ağın örneği olarak incelenmektedir.

Enron veri kümesi Enron şirketi ile ilgili yasal bir araştırma sırasında halka açılmıştır. Bu veri yapısının tam olarak 148 kullanıcıya ait 520.000 mesaj içermekte olduğu bilinmektedir. Kullanıcıların bu şirkette farklı bölümlerde ve farklı görevlerde çalıştığı düşünüldüğünde kullanıcılara ait mesajlarında farklılıkların kullanıcıların ilgilendiği bölüm ile ilgili bilgi vereceği düşünülmektedir. Aynı iş tanımı altındaki kişilerin mesaj içerikleri birbirine yakın olabilmektedir. Bu tür varsayımlar, bu yapının genellikle sosyal ağ analizi, link analizi ve veri analizi için kullanılmasını sağlamaktadır.

Bu çalışma için 148 kullanıcı arasında iletilen 35.359 e-posta seçilmiştir. Diğer e-postaların şirket dışındaki kişileri iletildiği gözlenmiş ve o e-postalar yapıdan çıkarılmıştır. Bu yapı 8.266.887 kelime sayısına sahiptir ve 368.690 farklı kelimeden oluşur.

### 4.2. Deneyler

Model için topluluk sayısı 10, konu sayısı ise 25 seçilmiştir. 1000 tekrarlardan sonra topluluk konu dağılımı  $\theta$ , yazar-alıcı-topluluk dağılımı  $\gamma$  ve konu-kelime dağılımı  $\phi$  elde edilmiştir.  $\alpha$ ,  $\beta$  ve  $\mu$  parametreleri sırasıyla  $50/C$ , 0.01,  $50/T$  olarak ayarlanmıştır [15].

Bu bölümde, topluluk, konu, bağlantı ve çalışma süresi analizi yapılmıştır.

#### 4.2.1. Topluluk ve Konu Analizi

Bu bölümde, ağ üyelerinin konuştuğu konular ile ilgili bilgi verilecek ve üretilen topluluklar sunulacaktır. YATK modeli, ağa ait içerik ve bağlantı bilgisini kullanarak toplulukları oluşturmaktadır. Bu nedenle, modelin diğer model ile hem konu açısından hem de topolojik olarak karşılaştırılması gerekmektedir. KL yöntemi ile model içerik olarak, NOB ve Modülerlik yöntemi ile de topolojik olarak karşılaştırılması yapılmıştır.

YATK model, Enron yapısına ait e-postaların kelime analizini yaparak belli sayıda konu başlığı altında yapının özetini çıkartmaktadır. Üretilen konular, tahsis edildikleri topluluk hakkında bilgi vermektedir. Modelin elde ettiği bazı konular Tablo 4.1’de gösterilmektedir.

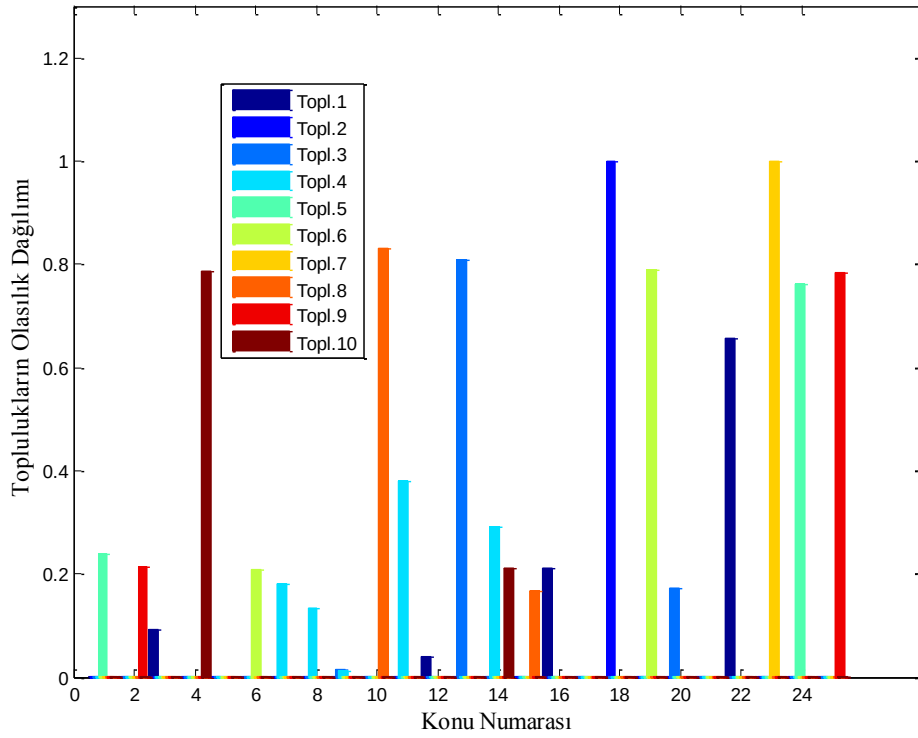
Tablo 4.1: Enron yapısına ait üretilen konular.

| Konu 1     |       | Konu 8       |       | Konu 22     |       | Konu 23  |       |
|------------|-------|--------------|-------|-------------|-------|----------|-------|
| contract   | 0.005 | million      | 0.006 | sager       | 0.009 | enron    | 0.133 |
| project    | 0.005 | finance      | 0.005 | enronxgate  | 0.009 | data     | 0.013 |
| question   | 0.005 | employees    | 0.003 | ect         | 0.008 | database | 0.013 |
| plan       | 0.004 | director     | 0.003 | transaction | 0.005 | id       | 0.010 |
| notes      | 0.004 | organization | 0.002 | energy      | 0.004 | error    | 0.010 |
| service    | 0.004 | corporate    | 0.002 | regulatory  | 0.004 | iso      | 0.010 |
| order      | 0.003 | bank         | 0.002 | power       | 0.003 | detected | 0.007 |
| commercial | 0.003 | profit       | 0.002 | provide     | 0.002 | bit      | 0.007 |
| cost       | 0.002 | valuation    | 0.002 | transfer    | 0.002 | text     | 0.007 |
| gas        | 0.002 | addition     | 0.002 | electricity | 0.002 | perform  | 0.007 |

Tablo 4.1’de 1. konunun ofis ve ticari işler ile ilgili olduğu sahip olduğu kelime dağılımından anlaşılmaktadır. Bu konuyu konuşan Enron çalışanlarının daha çok pazarlama işi ile ilgili oldukları düşünülebilir. 8. konunun şirketin finans işleri ile, 22. konunun enerji ve mühendislik işleri ile ve 13. konunun daha çok sistem ve veri tabanı ile ilgili olduğu görülmektedir. Bu konu başlıkları, bu konular ile ilgili yazışan kişilerin iş görevi hakkında tahmin yürütmemizi sağlayabilmektedir. Anlamsal benzerlik taşıyan kelimeler aynı konu başlığı altına tahsis edildiği görülür.

Şekil 4.1’de topluluk-konu dağılımı gösterilmektedir.  $y$  eksenindeki değer belli bir topluluğa tahsis edilmiş konu olasılığını,  $x$  eksenindeki değer ise konuları

göstermektedir. Konuya ait her bir çubuğun rengi de tahsis edildiği topluluğa göre değişmektedir. Her konu diğer konular ile karşılaştırıldığında yüksek bir olasılık ile bir topluluğa tahsis edilmiştir. Örneğin, 18. konunun 2. topluluğa 0.9'dan büyük bir olasılık ile atandığı görülmektedir. Bu durum toplulukların içerik olarak iyi ayrıştığını gösterebilir. Çok genel olmamakla beraber, bir konu birden fazla topluluğa tahsis edilebilir. 4. topluluğa 7., 8., 11. ve 14. konuları tahsis olmuştur. 1. konu haricinde diğer konuların tahsis edilme olasılıkları düşüktür. 22. konu 1. topluluğa 0.68 olasılık değeri ile atanmıştır. Bu durumda 1. topluluğa tahsis edilmiş kullanıcıların ve yazar –alıcı çiftlerinin daha çok 22. konu ile ilgili konuştukları sonucuna varılır.



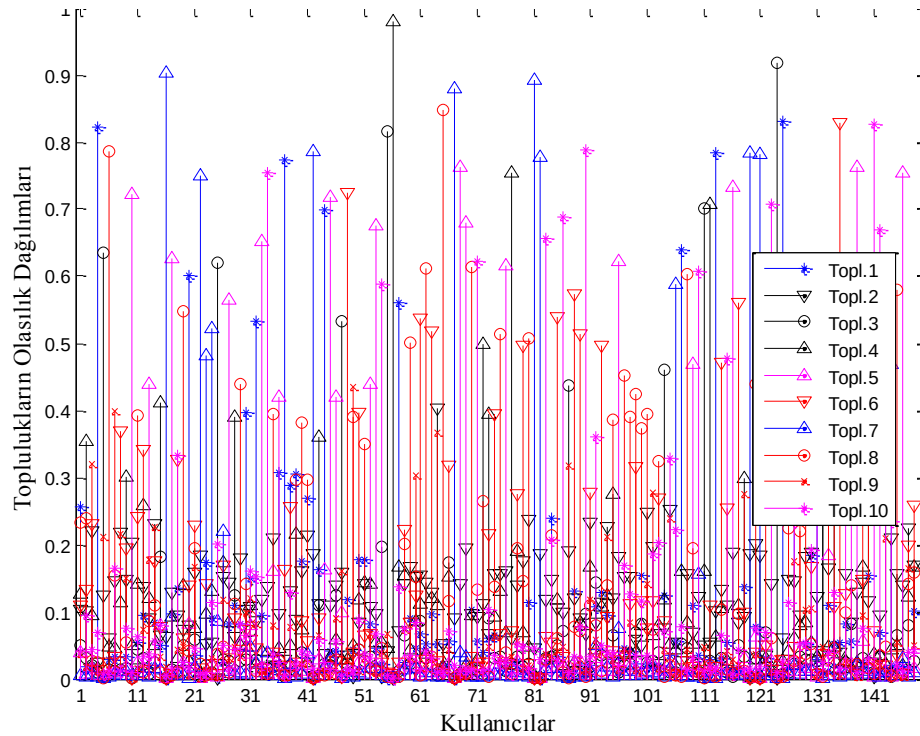
Şekil 4.1: Topluluk-konu dağılımı.

YATK ve TK modelinin topluluk-konu dağılımları KL metoduna göre analiz edildiğinde, YATK modelinin ürettiği topluluklar arasındaki KL farkının daha yüksek olduğu görülmektedir. Tablo 4.2’de iki modele ait KL sonucu verilmiştir.

Tablo 4.2: YATK ve TK modele ait ortalama KL deęerleri.

| Topluluk Numarası |      | 2       | 3       | 7       | 8       |
|-------------------|------|---------|---------|---------|---------|
| Ortalama KL       | YATK | 0,08639 | 0,08639 | 0,07474 | 0,07335 |
|                   | TK   | 0,05501 | 0,05408 | 0,05564 | 0,04898 |

Yüksek KL farkı topluluklar arasındaki konu benzerliğinin daha az olduğunu gösterir. Bu durum YATK modelinin ürettiğı toplulukların, TK ve dięer modellerle kıyaslandığında konu dağılımı bakımından daha iyi ayrıştığını göstermektedir. YATK modelinin ürettiğı topluluklardan 2., 3., 7. ve 8. Topluluklar seçilmiş ve KL deęerleri hesaplanmıştır. Örneğin, 2. Topluluğun tüm topluluklara olan uzaklıkları bulunmuş ve ortalaması alınarak ortalama KL deęeri hesaplanmıştır. Bu 4 topluluğun her biri için görüldüğü üzere yüksek KL deęerleri bulunmuştur. Bunun nedeni, her bir konunun yüksek bir olasılık ile topluluklara tahsis edilmesidir. TK modelinden daha yüksek sonuçlar bulunmasının neden modele yazar ve alıcı bilgisinin eklenmesidir.



Şekil 4.2: Kullanıcı-topluluk dağılımı.

YATK modelinin ürettiğı topluluklara belli olasılıklar ile tahsis edilmiş kullanıcılar Şekil 4.2’de verilmektedir. Ortak ilgi alanlarına sahip, dięer bir deyişle

konu dağılımları benzer kullanıcılar aynı topluluk altında kümelenmiştir. Enron sosyal ağına ait kullanıcıların topluluk dağılımı incelenerek, her bir kullanıcı hakkında bilgi edinilir.

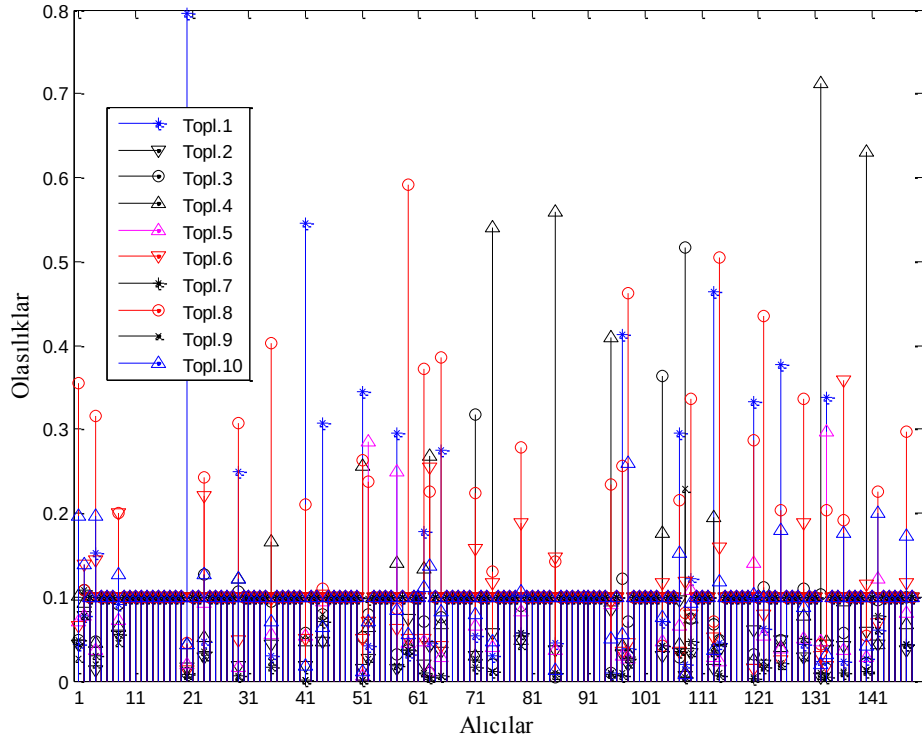
#### 4.2.2. Bağlantı Analizi

YATK modeli, toplulukların sosyal ağın topolojik yapısına benzer bir şekilde oluşturulmasını sağlar. Böylece oluşan topluluklar analiz edildiğinde ağın bağlantı yapısı da analiz edilebilmektedir. YATK modeline ait topluluklar, diğer yöntemlerle kıyaslandığında gerçek sosyal ağa daha çok benzemektedir. Bu karşılaştırma NOB yöntemi ile yapılmış ve sonuçları Tablo 4.3'te gösterilmiştir. YATK modeline ait 1. topluluk 0.1516 kadar benzerlik gösterirken, CT modeldeki benzerlik değeri oldukça düşük seyretmektedir.

Tablo 4.3: YATK ve TK model için NOB değerleri.

| Topluluk Numarası |      | 2      | 4      | 5      | 8       |
|-------------------|------|--------|--------|--------|---------|
| NOB               | YATK | 0.1516 | 0.3396 | 0.3411 | 0.3974  |
|                   | TK   | 0.0191 | 0.0390 | 0.0007 | 0.00006 |

YATK modeli mesaj göndericisi ve alıcısı arasındaki mesajlaşmalara göre yazar-alıcı çiftlerini topluluklara tahsis etmektedir. Eğer bir kullanıcı başka bir kullanıcıya mesaj gönderir ise aralarında bir ilişki başlatılmış olunur. Bu bağlantının hangi konular üzerinde devam ettiğinin analizi sosyal ağların analizi için önemlidir. Sosyal ağlarda bu bağlantılar belli ağırlıklara sahiptir. Genellikle bu ağırlık değerleri, iki kişi arasındaki mesajlaşma sayısı ile belirlenir. YATK modeli bağlantıları topluluklara tahsis edilme olasılıkları ile tekrar üretir. Böylece, bir topluluğa tahsis edilmiş iki kişinin arasındaki bağlantının ne kadar güçlü olduğunun analizi yapılabilmektedir. Yazar-alıcı çiftlerinin hangi topluluklar altında aktif olduğu izlenerek hangi konular hakkında konuştukları öğrenilebilmektedir.



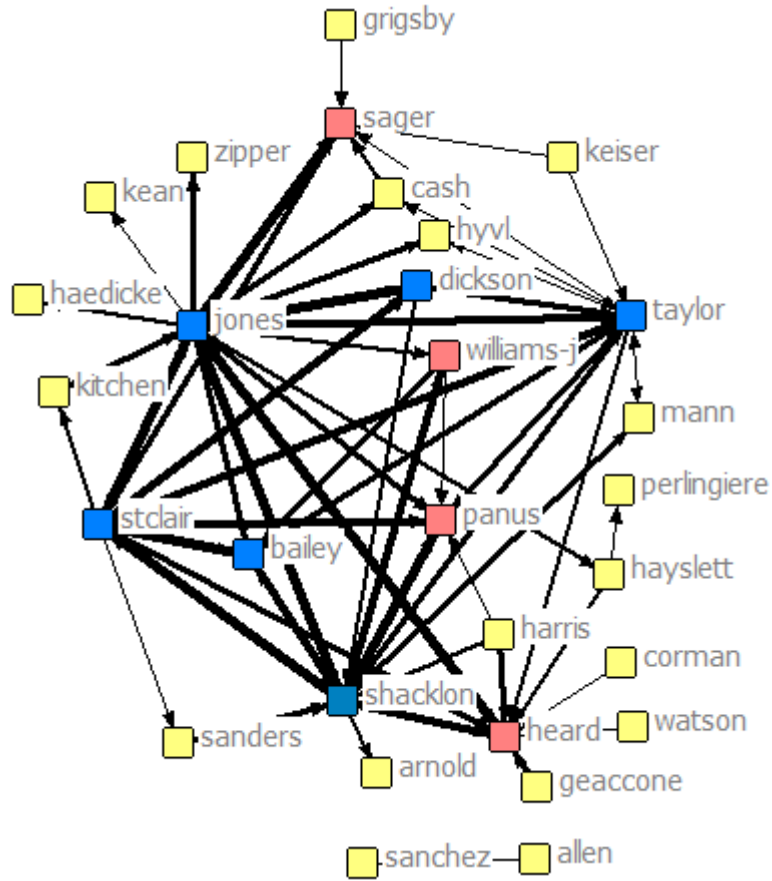
Şekil 4.3: 1. yazar için alıcı-topluluk dağılımı.

Şekil 4.3, Enron ağına ait 1.yazar için alıcı dağılımını göstermektedir. Şeklin  $x$  ekseninde 1. yazara ait alıcılar,  $y$  ekseninde 1. yazara ait alıcıların belli bir topluluğa tahsis edilme olasılıkları verilmiştir. Yazar-alıcı çiftlerinin tahsis edildiği topluluklar farklı şekil ve renklerle gösterilmiştir. 1. yazarın ve alıcıların daha çok 8. topluluğa atandığı görülür. Ayrıca 1. ve 4. topluluğa da atanma sayısı diğer topluluklara göre fazladır. 1. yazarın 8., 1. ve 4. topluluklara ait konularla ilgili alıcılar ile mesajlaştığı anlaşılabilir. Yazar-alıcı çiftlerinin hangi topluluklarda daha aktif olduğu bu sayede gözlemlenebilir ve bağlantı analizi yapılabilir.

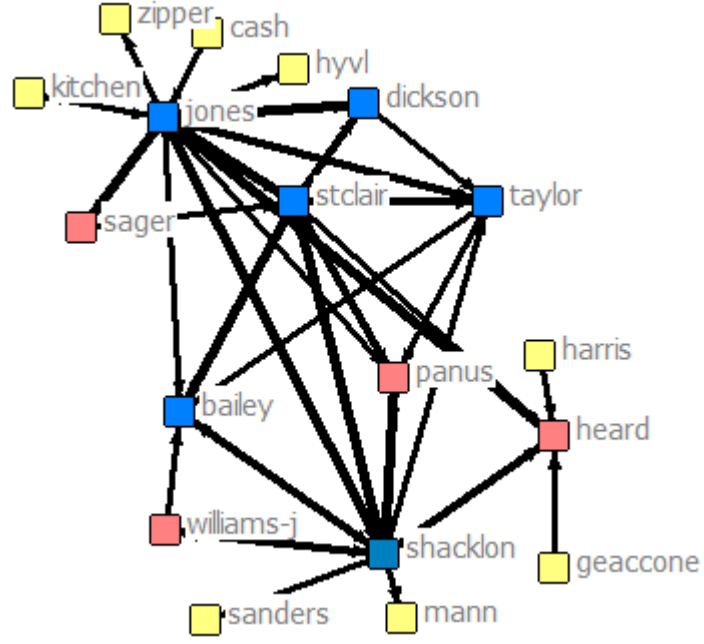
Yazar-alıcı çiftlerin topluluklar altında görüntüleri Şekil 4.4’de ve 4.5’de verilmektedir. Şekillerden anlaşıldığı üzere, her bir topluluk sosyal ağın bir parçası gibi görünmektedir. Çünkü kişi bilgisi yanında ağdaki bağlantıların ve bağlantı ağırlıklarının analizi de yapılabilmektedir. Model, kullanıcıları bir topluluğa atarken, aynı zamanda ağdaki bağlantıları da topluluklara tahsis etmektedir. Bu durum, şüpheli görünen kişiler arasında konuşulan konuların deşifre edilmesi ve aynı şüpheli konuları konuşan kişilerin bir topluluk altında incelenmesi sağlanmaktadır.

Şekil 4.4, 4.5 ve 4.6’da 3. topluluğun farklı durumları görüntülenmiştir. Topluluklara ait düğüm ve bağlantı bilgisi açık bir şekilde görülmektedir. Topluluğa yüksek olasılıklar ile atanmış kişiler mavi ile gösterilmektedir. Topluluğu

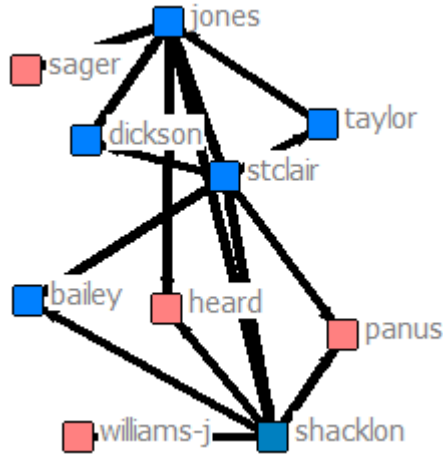
görsüntölemek için bir eşik değeri kullanılmıştır. Bu eşik değeriinin kullanılmasının sebebi, ağdaki üyelerin en az bir kere de olsa iletişime geçtiği düşünöldüğünde olasılıklara bağılı olarak bulunan bağlantı sayısı epeyce fazla olmaktadır. Bu yüzden topluluğu en yüksek olasılıkla temsil eden bağlantılar görsüntölenmiştir. Sırasıyla 0.5, 0.65 ve 0.9'dan eşit ve büyük bir olasılıkla topluluğa atanmış yazar-alıcı çiftleri görsüntölenmektedir. Düğümler arasındaki bağlantı çizgileri, Topluluk 3'e tahsis olma olasılığına göre farklı kalınlıkta gösterilmiştir. Diğerlerine göre daha büyük olan bağlantılar daha kalın çizilmiştir. Şekilde göröldüğü üzere "jones" ve "stclair" arasındaki bağlantının değeri yüksektir. Bu nedenle aralarındaki çizgi kalın çizilmiştir. Topluluğa en yüksek olasılık değeri ile atanmış kullanıcılar sırasıyla pembe, mavi ve sarı renkle gösterilmektedir. "stclair", "jones", "shacklon" ve "taylor" kullanıcıların yüksek ağırlıklı bağlantılara sahip oldukları görölmektedir.



Şekil 4.4: 3. topluluğa ait ağ görsüntüsü, eşikdeğeri  $\geq 0,5$ .



Şekil 4.5: 3. topluluğa ait ağ görüntüsü, eşikdeğeri  $\geq 0,65$ .



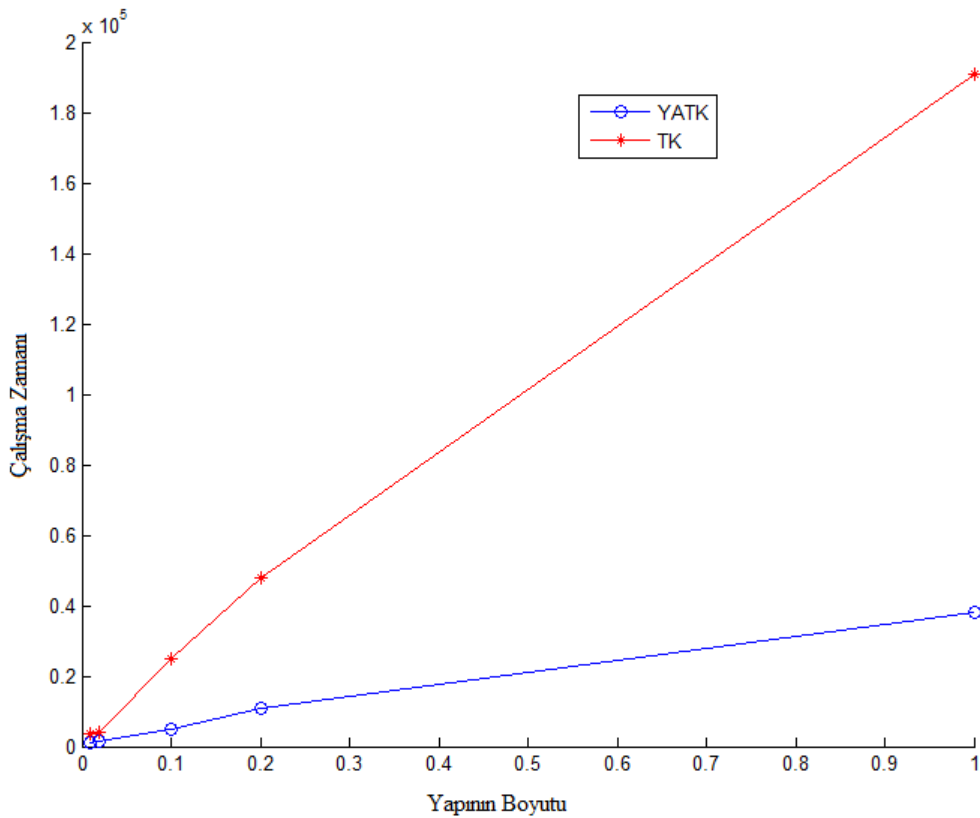
Şekil 4.6: 3. topluluğa ait ağ görüntüsü, eşikdeğeri  $\geq 0,9$ .

#### 4.2.3. Çalışma Süresi Analizi

Sosyal ağlar milyonlarca düğüm ve kenarlara sahip olabilmektedir. Sosyal ağların popülaritesinin artması ile bu sayılar da hızlı bir artış görülmektedir. Sosyal ağlarda konu açısından analiz yapan konu modelleri için her bir sosyal ağ üyesinin sahip olduğu doküman sayısı da çok önemlidir. Çünkü konu modelleri yüksek hesaplama karmaşıklığına konusunda sıkıntı çeken Gibbs örnekleme algoritmasını

kullanılmaktadır. Bu sebeple, çalışma zamanı topluluk tanıma modelleri için büyük bir problemdir.

TK modelinde, her bir dokümana ait yazarlar tek tek işleme alınmaktadır. Bir sosyal ağda bir üyenin gönderdiği ve aldığı mesaj o kişi hakkında bilgi verir. Bu nedenle her bir e-posta için hem gönderici hem de alıcı için Gibbs örnekleme çalışacaktır.  $I$  sayıda tekrarlama,  $T$  sayıda konu,  $C$  sayıda topluluk için  $n$  sayıda kelimeye,  $k$  sayıda yazara ve  $m$  sayıda alıcıya sahip bir e-postaya ( $E = 1$ ) ait Gibbs örneklemenin en kötü karmaşıklık değeri  $O(I * E * (K + M) * N * C * T)$ 'dir. YATK modelinde bir e-postaya ait yazar ve alıcılar ayrı incelenmektedir. Bir e-posta için bir yazar incelenirken, o e-postaya ait alıcılar kümesinden bir alıcı rastgele seçilerek o yazar ile ilişkilendirilmektedir. Bu nedenle, bir e-posta için sadece yazar sayısı esas alınır. E-posta sayısının 35.359 olduğu Enron yapısı için, YATK modeline ait karmaşıklık değeri  $O(I * E * K * N * C * T)$  olmaktadır.



Şekil 4.7: YATK ve TK modeline ait çalışma süreleri.

Çalışma zamanı karşılaştırılması Şekil 4.7'de verilmektedir. Doküman sayısına göre iki modelin çalışma zamanları görülmektedir. Çalışma zamanı

karşılaştırmasında YATK modelin farklı veri boyutlarında daha kısa sürede sonuç ürettiği görülmektedir.

## 5. SONUÇ VE ÖNERİLER

Bu çalışmada, sosyal ağları analiz etmek için sosyal topluluk-tanırma yöntemi olan YATK model geliştirilmiştir. Sosyal ağlardaki dokümanları konusal olarak analiz etmenin yanı sıra, ağın topolojisini muhafaza ederek, konu benzerliklerine göre ağın bağlantı yapısını benzer konu dağılımına sahip olacak şekilde topluluklara ayırmaktadır. Ağ yapısının topolojik olarak bir parçası olacak şekilde topluluklar üretilmiştir. Konu içerik bilgisi yanında bağlantı bilgisine göre toplulukların oluşturulması ağın bağlantı analizinin yapılabilmesine olanak sağlamıştır. Konu modelleri temelli topluluk tanırma yöntemleri topolojik olarak ayrı alt topluluklar oluşturmaktadır. Geliştirilen YATK model ile sosyal ağ hem topolojik hem de konusal olarak topluluklara ayrıştırılmaya çalışılmaktadır.

Geliştirilen yöntem birçok uygulama için kullanım alanı sağlamaktadır. Şüpheli ağ analizinin yanı sıra, sosyal ağlarda bağlantı içerik analizi için kullanılabilir. YATK modelinin dinamik versiyonu yapılarak, sosyal ağlarda zamana göre değişen konular ve kişiler incelenebilir. Suçlu ağları konusunda daha etkin kullanabilmek için sosyal ağlarda kullanılan suçla ilişkili kelimelere ait dinamik bir sözlük oluşturulabilir.

## KAYNAKLAR

- [1] Steyvers M., Smyth P., Rosen-Zvi M., Griffiths T., (2004), "Probabilistic author-topic models for information discovery", In Proceedings of The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 306-315, Seattle, WA, USA, 22-25 August.
- [2] Blei D. B., Ng A. Y., Jordan M. I., (2003), "Latent dirichlet allocation", The Journal of Machine Learning Research, 3, 993-1022.
- [3] Xia Z., (2008), "Fighting criminals: Adaptive inferring and choosing the next investigative objects in the criminal network", Knowledge-Based Systems, 21, 434-442.
- [4] L'Huillier G., Ríos S. A., Alvarez H., Aguilera F., (2010), "Topic-based social network analysis for virtual communities of interests in the Dark Web", In ACM SIGKDD Workshop on Intelligence and Security Informatics, 9, Washington, DC, USA, 25-28 July.
- [5] Zhou D., Manavoglu E., Li J., Giles C. L., Zha H., (2006), "Probabilistic models for discovering e-communities", In Proceedings of the 15th International Conference on World Wide Web, 173-182, Edinburg, Scotland, UK, 23-26 May.
- [6] Al-Zaidy R., Fung C. M., Youssef A., (2011), "Towards discovering criminal communities from textual data", In Proceedings of the ACM Symposium on Applied Computing, 172-177, Taichung, Taiwan, 21-24 March.
- [7] Rosen-Zvi M., Chemudugunta C., Griffiths T., Smyth P., Steyvers M., (2008), "Learning author-topic models from text corpora", ACM Transactions Information Systems, 28 (1), 4.
- [8] Baumgartner K., Ferrari S., Palermo G., (2008), "Constructing Bayesian networks for criminal profiling from limited data", Knowledge-Based Systems, 21, 563-572.
- [9] Xu J., Chen H., (2005), "Criminal network analysis and visualization", Communication, 48 (6), 100-107.
- [10] Xu J., Chen H., (2004), "Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks", Decision Support Systems, 38 (3), 473-487.
- [11] Newman M., (2004), "Detecting community structure in networks", The European Physical Journal B-Condensed Matter and Complex Systems, 38 (2), 321-330.

- [12] Girvan M., Newman M., (2002), "Community structure in social and biological networks", *Proceedings of the National Academy of Sciences*, 99 (12), 7821-7826, Washington, DC, USA, 11 June.
- [13] McCallum A., Andrés C., Xuerui W., (2005), "The author-recipient-topic model for topic and role discovery in social networks with the application to Enron and academic email", *Workshop on Link Analysis, Counterterrorism and Security*, 33-44, Newport Beach, California, USA, 23 April.
- [14] Pathak N., DeLong C., Banerjee A., Erickson K., (2008), "Social topic models for community extraction", In *The 2nd SNA-KDD Workshop*, 8, Las Vegas, Nevada, USA, 24-27 August.
- [15] Li D., Ding Y., Shuai X., Bollen J., Tang J., Chen S., Zhu J., Rocha G., (2012), "Adding community and dynamic to topic models", *Journal of Informetrics*, 6 (2), 237-253.
- [16] Euler L., (1741), "Solutio problematis ad geometriam situs pertinentis", *Commentarii academiae scientiarum imperialis Petropolitanae*, 8, 128-140.
- [17] Web 1, (2013), [http://tr.wikipedia.org/Königsberg'in\\_yedi\\_köprüsü](http://tr.wikipedia.org/Königsberg'in_yedi_köprüsü), (Erişim Tarihi: 14/07/2013).
- [18] Klerks P., (2003), "The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators?", *Recent developments in the Netherlands Organised Crime: Perspectives on Global Security*, 97-113.
- [19] Harper W. R., Harris D. H., (1975), "The application of link analysis to police intelligence", *The Journal of the Human Factors*, 17, 157-164.
- [20] Krebs V. E., (2001), "Mapping networks of terrorist cells", *Connections*, 24, 43-52.
- [21] Chau M., Xu J. J., Chen H., (2002), "Extracting Meaningful Entities from Police Narrative Reports", *Proceedings National Conference Digital Government Research*, 271-275, Los Angeles, CA, USA, 19-22 May.
- [22] Hauck R. V., Atabakhsb H., Ongvasith P., Gupta H., Chen H., (2002), "Using Coplink to analyze criminal-justice data", *Computer*, 35 (3), 30-37.
- [23] Doreian P., Stokman F. N., (1997), "The dynamics and evolution of social networks", In *Evolution of Social Networks*, 1-17, New York, NY, USA.
- [24] Xu J., Marshall B., Kaza S., Chen H., (2004), "Analyzing and visualizing criminal network dynamics: A case study", In *Intelligence and Security Informatics*, Springer, 359-377.

- [25] Freeman C., (1979), "Centrality in social networks: Conceptual clarification", *Social Networks* , 1, 215-240.
- [26] Alpaydin E., (2004), "Introduction to machine learning", 1th Edition, The MIT Press.
- [27] Duardo-Sanchez A., (2010), "Study of criminal law networks with Markov-probability centralities", *Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks*, 205-212, Bentham, Kerala, India.
- [28] Leenders R., (1996), "Evolution of friendship and best friendship choices", *The Journal of Mathematical Sociology*, 21, 133-148.
- [29] Hummon N. P., (2000), "Utility and dynamic social networks", *Social Networks*, 22, 221-249.
- [30] Bayes T., Price M., (1763), "An Essay towards Solving a Problem in the Doctrine of Chances", *Philosophical Transactions (1683-1775)*, 370-418.
- [31] Salton G., McGill M. J., (1986), "Introduction to Modern Information Retrieval", McGraw-Hill, New York, NY, USA.
- [32] Papadimitriou C. H., Tamaki H., Raghavan P., Vempala S., (1998), "Latent semantic indexing: a probabilistic analysis", In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 159-168, Seattle, Washington, USA, 1-3 June.
- [33] Hofmann T., (1999), "Probabilistic latent semantic indexing", In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-57, Berkeley, CA, USA, 15-19 August.
- [34] Geman S., Geman D., (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.
- [35] Griffiths T., Steyvers M., (2004), "Finding scientific topics", *Proceedings of the National Academy of Sciences of the United States of America*, 101, 5228-5235, USA, 6 April.
- [36] Huang A., (2008), "Similarity measures for text document clustering", *Proceedings of New Zealand Computer Science Research Student Conference*, 49-56, Christchurch, New Zealand, 14-18 April.
- [37] MacKay D. J., (2003), "Information Theory, Inference, and Learning Algorithms", 1th Edition, Cambridge University Press.
- [38] Newman M. E., Mark E. J., Girvan M., (2004), "Finding and evaluating community structure in networks", *Physical Review E*, 69 (2), 026113.

- [39] Klimt B., Yiming Y., (2004), “The Enron Corpus: A new dataset for email classification research”, In Machine Learning: ECML, 217-226.

## ÖZGEÇMİŞ

Arzu KAKIŞIM, 1985 yılında Çankırı’da doğmuştur. İlköğretim ve lise eğitimini Ankara’da tamamlayıp 2009 yılında Karadeniz Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü’nden lisans mezunu olmuştur. 2010 yılında Gebze Yüksek teknoloji Enstitüsü Mühendislik ve Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı’nda yüksek lisans eğitimine başlamıştır. Ayrıca 2011 yılından itibaren Gebze Yüksek Teknoloji Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı’nda araştırma görevlisi olarak çalışmaktadır.