Büyük Veride Kişi Mahremiyetinin Korunması

Can EYÜPOĞLU[1]*, Muhammed Ali AYDIN[2], Ahmet SERTBAŞ[2], Abdül Halim ZAİM[3], Onur ÖNEŞ[4]

[1]Bilgisayar Mühendisliği Bölümü, İstanbul Ticaret Üniversitesi, İstanbul, Türkiye
[2]Bilgisayar Mühendisliği Bölümü, İstanbul Üniversitesi, İstanbul, Türkiye
[3]Bilişim Teknolojileri Uygulama ve Araştırma Merkezi, İstanbul Ticaret Üniversitesi, İstanbul, Türkiye
[4]Bilişim ve İletişim Teknolojileri, Turkcell Teknoloji, İstanbul, Türkiye
ceyupoglu@ticaret.edu.tr, aydinali@istanbul.edu.tr, asertbas@istanbul.edu.tr, azaim@ticaret.edu.tr, onur.ones@turkcell.com.tr
(Geliş/Received:26.09.2016; Kabul/Accepted:03.02.2017)
DOI: 10.17671/gazibtd.309301

***Özet*—**Büyük verinin ortaya çıkışı bilgi güvenliği ve klasik güvenlik tedbirleri için kullanılan koruma modelleri için yeni zorluklara neden olmaktadır. Bu çalışmada büyük veri güvenliği ve büyük veride kişi mahremiyetinin korunmasına yönelik literatürde var olan çalışmalar özetlenmiştir. Buna ek olarak büyük veri kaynaklarının neler olduğu, büyük veri sistemlerini korumak için gerekli olan araçlar ve bu sistemlerin güvenliğinin sağlanmasında karşılaşılan zorluklar açıklanmıştır.

***Anahtar Kelimeler*—**Büyük veri, büyük veri güvenliği, kişi mahremiyeti

Preserving Individual Privacy in Big Data

***Abstract*—**The emergence of big data causes new difficulties for protection models used for information security and conventional security precautions. In this study, existing studies in the literature related to big data security and preserving individual privacy in big data are summarized. Besides, big data sources, necessary means to protect big data systems and the difficulties in ensuring the security of these systems are explained.

***Keywords*—**Big data, big data security, individual privacy

1. GİRİŞ (INTRODUCTION)

Büyük veri; sosyal medya paylaşımları, GSM operatörlerinden elde edilen arama kayıtları, fotoğraf, video, blog ve log dosyaları gibi farklı kaynaklardan elde edilen verilerin anlamlı ve işlenebilir hale dönüştürülmüş biçimidir. Şu anda yapısal veriler ilişkisel veritabanlarında tutulmaktadır ve birçok şirket günlük olarak terabaytlarca veriyi yönetmek zorundadır. Geleneksel altyapı bu tür veriler üzerinde işlemler gerçekleştirmek ve bunların analitik verilerini oluşturmak için eksik kalmaktadır. İlişkisel veritabanlarını kullanan şirketler veri madenciliği yöntemleri ile elde ettikleri verileri kullanarak karar almaktadır. Fakat günümüzde bu tip bir çözüm yeterli değildir. Firmalar yapısal olmayan büyük veriyi oluşturmak, analiz etmek ve saklamak zorundadır. Klasik veritabanları sürekli artan bu boyuttaki veriyi tutmakta başarılı olamamaktadır. Ayrıca büyük verinin ortaya çıkışı kişi mahremiyetinin korunmasına yönelik yeni zorluklara neden olmaktadır. Geleneksel veritabanları için birçok başarılı kimliksizleştirme ve veri güvenliği çözümü bulunmaktadır. Ancak büyük veri üzerinde geliştirilmeye başlanan bu tip çözümler daha emekleme aşamasındadır.

Organizasyonlar güvenliği ve mahremiyeti sağlamak için çeşitli kimliksizleştirme (de-identification) yöntemleri kullanmaktadır. Güvenliği ve mahremiyeti garanti altına almak için en yaygın çözüm sözlü ve yazılı taahhütlerdir. Fakat bu çözümün başarısız olduğu görülmüştür. Şifreler, denetimli erişim ve iki-faktörlü kimlik doğrulama; veri dinamik ve dağıtık veri sistemlerinde paylaşıldığı ve bir araya getirildiğinde güvenliği ve gizliliği sağlamak için düzenli olarak kullanılan teknik çözümlerdir ancak düşük seviyelidir. Daha gelişmiş teknik çözüm ise kriptografidir. AES ve RSA tanınmış şifreleme algoritmalarıdır. Son günlerdeki açıklamalar NSA (National Security Administration)'nın mevcut Internet şifreleme algoritmalarını kırmanın yollarını bulduğunu göstermektedir [1, 2].

Güvenlik duvarları (firewalls), güvenli soket katmanı, taşıma katmanı güvenliği gibi sanal engeller veriye erişimi sınırlamak için tasarlanmıştır. Bu teknolojilerin hepsi kırılabilmektedir. Bu sebeple sürekli olarak gözlemlenmelidir ve gerektiğinde onarılmalıdır. İzleme, gözlemleme veya denetleme yazılımları güvenliğe uygunluğu garanti altına almak için bireysel kullanıcıların veri akış ve ağ erişim geçmişlerini temin etmek için tasarlanmıştır. Bu teknolojinin kısıtlaması büyük ölçekli

veya dağıtık veri sistemlerinin gerçekleştirilmesinin zor ve maliyetli olmasıdır. Çünkü bulguların okunması ve yorumlanması gerekmektedir. Ayrıca yazılım veriyi korumadan ziyade bireysel davranışları gözlemlemede kullanılabilir. Sonuç olarak geleneksel kimliksizleştirme teknikleri yaygın kullanımları nedeniyle büyük veri çağında uygulanabilir değildir. Büyük veri güvenliği ve gizliliğini sağlama görevleri bilgi arttıkça daha zor olmaktadır. Bilgisayar uzmanları isimsizleştirilmiş (anonymized) verinin sıklıkla tekrardan tespit edildiğini ve belirli bir bireye dayandırıldığını sürekli olarak göstermektedir [1, 3].

Sosyal ağ büyük veride önemli bir veri kaynağıdır. Fakat bu veri büyük oranda kullanıcıların özel verisini içermektedir. Sosyal ağ anonimlik koruması kullanıcı ID'si için anonimliği gerçekleştirmelidir. Sosyal ağ anonimlik korumasının en önemli sorunu saldırganın anonim bir kullanıcı olabilmesidir. Bu anonim kullanıcı, kullanıcıların halka açık olan bilgileri ve kullanıcılar arasındaki bağlantıları kullanarak sonuç çıkarabilir [1]. Örnek olarak çeşitli bağlantı tahmin algoritmaları [4-8]'de öne sürülmüştür.

Çalışmanın ikinci bölümünde büyük veri sistemlerini korumak için gerekli olan araçlardan bahsedilecektir. Üçüncü bölümde büyük veri sistemlerinin güvenliğinin sağlanmasındaki zorluklara değinilecektir. Dördüncü bölümde büyük veri kaynaklarına yer verilecektir. Beşinci bölümde ise büyük veri güvenliği için geliştirilen yöntemlerden bahsedilecek ve altıncı bölümde sonuç verilerek makale sonlandırılacaktır.

2. BÜYÜK VERİ SİSTEMLERİNİ KORUMAK İÇİN GEREKLİ OLAN ARAÇLAR (NECESSARY MEANS FOR PROTECTION OF BIG DATA SYSTEMS)

Büyük veri bilgi güvenliği ve geleneksel güvenlik tedbirleri için benimsenen koruma ideolojisi için yeni tehditlere yol açmaktadır. Büyük veri güvenliği konularını çalışan bir grup olan CSA (Cloud Security Alliance, Bulut Güvenlik Antlaşması) son olarak büyük veri sistemlerini korumak için gerekli olan araçları listeleyen bir doküman hazırlamıştır [9, 10]:

1. Dağıtık (distributed) programlama çerçeveleri (framework) için güvenilir hesaplamalar,
2. İlişkisel olmayan veri depolama alanları için güvenlik uygulamaları,
3. Güvenilir veri depolama ve işlem logları,
4. Uç nokta giriş onaylama/filtreleme,
5. Gerçek zamanlı güvenlik görüntüleme,
6. Ölçeklenebilir ve birleştirilebilir gizlilik korumalı veri madenciliği ve matematiksel analiz,
7. Kriptografik zorunlu veri merkezli güvenlik,
8. Granüler erişim kontrolü,
9. Granüler hesap denetimleri,
10. Veri kaynağı (provenance).

CSA bu araçları dört gruba bölmüştür: altyapı güvenliği, veri koruma, veri yönetimi ve reaktif güvenlik.

Hadoop dağıtık programlama çerçevesine örnek olarak verilebilir. Hadoop Java ile yazılmış açık kaynak kodlu bir çerçevedir ve sıradan sunucular üzerinde büyük veri işlemeye yarar. HDFS (Hadoop Distributed File System, Hadoop Dağıtık Dosya Sistemi) birden çok bilgisayar üzerinde dağıtık olarak çalışan bir dosya sistemidir ve makineler üzerindeki dosya sistemlerini birbirine bağlayarak tek bir dosya sistemi gibi gözükmesini sağlar. İlişkisel olmayan veri depolama alanları yani büyük veri ortamları için geliştirilecek güvenlik uygulamaları, bu uygulamalarda yer alacak olan güvenilir veri depolama ve güvenilir işlem logları büyük öneme sahiptir. Ayrıca bu sistemlerde uç nokta girişlerin onaylanması, filtrelenmesi ve gerçek zamanlı olarak sistemin görüntülenmesi gerekmektedir. Veri madenciliği uygulamalarının ve matematiksel analizlerin ise kişi mahremiyeti göz önünde bulundurularak yapılması gerekmektedir. Bunlara ek olarak kriptografik uygulamaların yer aldığı merkezi bir güvenlik birimine ihtiyaç duyulmaktadır. Granüler olarak hesap erişimlerinin denetlenmesi, kontrolü ve veri kaynağının güvenilir olması büyük veri sistemlerini korumak için gerekli olan diğer araçlardır.

3. BÜYÜK VERİ SİSTEMLERİNİN GÜVENLİĞİNİN SAĞLANMASINDAKİ ZORLUKLAR (CHALLENGES IN PROVIDING SECURITY OF BIG DATA SYSTEMS)

Siber güvenlik problemleri için olan büyük veri analiz uygulamaları önemli derecede problemleri çözse de gerçek potansiyeli gerçekleştirmek için birkaç zorluğa değinmek gerekmektedir. Bu zorluklar şunlardır [9]:

1. Gizlilik (mahremiyet),
2. Büyük veri analizi APT (Advanced Persistent Threats, Gelişmiş Kalıcı Tehditler) algılama,
3. Yüksek performanslı kriptografi,
4. Güvenlik araştırması için büyük veri setleri,
5. Veri kaynağı problemi,
6. Güvenlik görselleştirme,
7. Nitelikli personel.

Büyük verinin ortaya çıkışı ile bilgi güvenliği ve geleneksel güvenlik tedbirleri için olan çözümler yetersiz kalmaktadır. Bu noktada kişilerin mahremiyeti büyük öneme sahiptir. Büyük veri sistemlerinde ortaya çıkabilecek olan kalıcı tehditlerin algılanması gerekmektedir ve gerekli noktalarda yüksek performanslı kriptografik uygulamalar kullanılmalıdır. Büyük veride veri kaynağının ve araştırmalarda kullanılacak olan veri setlerinin güvenilir olması çok önemlidir. Ayrıca büyük veri sistemlerini tasarlayan, geliştiren ve güvenlik problemlerini gözlemleyen kişilerin bilgi birikiminin yeterli düzeyde olması gerekmektedir.

Büyük veri için olan tehditler ve büyük veri altyapısı için düşünülmesi gerekenler şöyledir [11, 12]:

1. İşlemeyen verinin korunması,
2. Yönetimsel veri erişimi,
3. Yapılandırma ve yama (patch) yönetimi,
4. Uygulamaların ve düğümlerin kimlik doğrulamaları,
5. Denetim ve loglama,
6. İzleme, filtreleme ve engelleme,
7. API (Application Programming Interface, Uygulama Programlama Arayüzü) güvenliği.

Geliştirilecek olan büyük veri sistemlerinin altyapısında işlenmeyen yani sadece depolanan verinin korunması, verilere erişimin yönetilmesi ve var olan sistemin geliştirilmesi için eklenen yamaların yönetimi çok önemlidir. Bu sistemlerde uygulamaların ve düğümlerin kimlik doğrulamalarının yapılması, denetlenmesi ve log dosyalarının tutulması gerekmektedir. Ayrıca bu sistemler izleme, filtreleme, engelleme ve API güvenliği gibi mekanizmalara sahip olmalıdır.

4. BÜYÜK VERİ KAYNAKLARI (BIG DATA SOURCES)

Güvenlik analinizin daha güvenli olması için veri kaynakları göz önünde bulundurulmalıdır. Güvenlik analizleri (analytics) için olan büyük veri kaynakları şunlardır [11, 13]:

1. Sistem-tabanlı veri: IP konumları (locations), klavye yazma veya mouse tıklama akış desenleri vb.,
2. Mobil-tabanlı veri: GPS konumları, ağ konumları vb.,
3. Ağın fiziksel fazlalığının zamanı ve konumu,
4. Hareket (ulaşım) verisi: hareket desenleri, kaynakları, hedefleri vb.,
5. Dış yetkisiz kaynaklardan gelen veri,
6. Kimlik verisi: kullanıcı ismi ve şifresi vb.,
7. OTP (One Time Passwords): online erişim için kullanılan tek seferlik şifre,
8. Kimlik doğrulamada (authentication) kullanılan kullanılan dijital sertifikalar,
9. Biyometrik kimlik saptama (identification) verisi: parmak izi, iris, ses tanıma (speech recognition),
10. Sosyal medya verisi: Facebook, Google Drive, Twitter vb.

Veriyi daha güvenli ve verimli hala getirmek için dikkat edilmesi gereken kaynaklar şunlardır [11]:

1. Ağ trafiği,
2. Ağ kaynakları,
3. Kullanıcı kimlik bilgileri,
4. Ağ sunucuları.

5. BÜYÜK VERİ GÜVENLİĞİ İÇİN GELİŞTİRİLEN YÖNTEMLER (METHODS DEVELOPED FOR BIG DATA SECURITY)

Büyük verinin gizliliği en zorlayıcı faktörlerden biri haline gelmiştir. Veri setlerinin dramatik bir şekilde artmasıyla depolama ve analiz ihtiyacı da artmaktadır. Bu sebeple güvenlik de artırılmalıdır. Çünkü yetkisiz bir kişi belirli bir kişinin verisini indirebilir ve ona zarar verebilir. Saldırganları takip etmek için farklı gözetleme (surveillance) araçları ve takip cihazları kullanılmaktadır. Fakat veri arttıkça güvenlik de artırılmalıdır. Genel olarak kişisel verinin toplanması, depolanması, analizi ve kullanımı güvenliğin her seviyesinde günlük hayatımızın bir parçası olmuştur. Bu problem için bir çözüm de desen gizlemedir (pattern hiding). Bu çözümde veri yalnızca analistin çözebileceği bir desen içerisinde saklanır. Bu veri setleri ayrıca veri toplayıcısı tarafından da tanınamaz olmalıdır. Aksi takdirde veri yetkisiz taraflara sızdırılabilir. Güvenli dağıtımlı madencilik (secure distributive mining) ayrıca çok sayıda veri seti üzerinde de yapılabilir. Burada madencilikte kullanılacak veri farlı gruplar arasında yatay veya dikey olarak parçalara ayrılır ya da dağıtılır. Bölümlenmiş veri paylaşılamaz ve özel kalmalıdır. Ancak birleştirilen veri üzerindeki madenciliğin sonuçları katılımcılar (participants) arasında paylaşılır [11].

Verinin k-anonimliği veya kimliksizleştirilmesi kullanıcıların mahremiyetini tehlikeye atmadan web üzerinden verilerin paylaşılması yeteneğini ifade etmektedir. Diğer bir değişle k-anonimlik bilginin açığa çıkma riskini sınırlarken veri kullanımını maksimuma çıkaran bir güvenlik kavramıdır. Buradaki bilgi açık bir kimlik olmasa bile ilgili varlığın kimliğinin tespitine yol açmaktadır [14].

Büyük veri kullanımının artan başarısı ile birçok zorluk ortaya çıkmıştır. Araştırmacıların çözmeye çalıştığı ana problemler zamansızlık (timeless), ölçeklenebilirlik (scalability) ve gizliliktir. Gizlilik koruma şu anda oldukça aktif bir araştırma alanıdır. Bu konu ile ilgili olarak birçok çalışma ve konsept vardır. Bu konseptlerden biri de kimliksizleştirme tekniğidir. Kimliksizleştirme hassas bilgiyi bulma ve silmeden oluşan özel bir alandır. Kimliksizleştirme kriptografi ve veri madenciliği gibi teknikleri kullanarak bilginin yerinin değiştirebilir, şifreleyebilir ya da bilgiye bir gürültü ekleyebilir. Rahmani ve diğerleri tarafından yapılan çalışmada [14] CLONALG olarak bilinen özel bir bağışıklık sistemi kullanarak metin verilerini kimliksizleştiren yeni bir model ortaya konulmuştur. Bu yaklaşımda kimliksizleştirme (de-identification) ve tekrardan kimlik saptama (re-identification) olmak üzere iki ana aşama vardır. Kimliksizleştirme kullanıcıların verilerinin içerisinde yer alan kimliklerinin korunmasını sağlayan ana aşamadır. Bu aşama belirleyici (identifier) olarak düşünülen herhangi bir bilginin bulunması ve yerinin değiştirilmesi için kullanılmaktadır. Bu aşama da beş adımdan oluşmaktadır: tokenların ayırma (tokenization), kodlama (codification), tespit etme (detection), depolama

(storage) ve yer değiştirme (replacement). Tokenlarına ayırma adımında bir kelime dizisi içerisinde kimliksizleştirme yapmak için kullanıcıların bilgisini ve veriyi ayrıştırırken bir kelime paketi kullanılmaktadır. Kodlama adımında kimliklerdeki her bir kelime karakterlerine ayrıştırılır. Ardından her biri kendi ASCII koduyla değiştirilir. Böylece her bir kelime bir antijen (antigen) olarak düşünülür. Aynı işlem kimliksizleştirilecek verideki kelimelere de uygulanır. Bu verideki kelimeler antikor (antibodies) olarak düşünülür. Tespit etme aşaması mesafe metriğini kullanarak veriden kimlik saptayıcılarının (identifier) tespit edilmesine olanak sağlar. Mesafe metriği her bir antijen ve tüm antikorlar arasında hesaplanır. Depolama adımında bellekteki tanınan antikorlar depolanmaktadır. Doğal bağışıklık sistemi sahibiyle birlikte doğan bir başlangıç değerine sahiptir. Bu çalışmadaki sistemde bu başlangıç değeri kullanıcının yazıtında (inscription) oluşturulur. Böylece yazıt formülünde doldurulan bilgi antijenlere dönüştürülmektedir ve bellekte depolanmaktadır. Sonrasında yeni antikorlar tanınıncaya kadar, bellek bu antikorları ekleyerek güncellenir. Sonunda bu antikorlar antijen olarak düşünülen kelimelerin bazı varyasyonlarını temsil eder. Kelimelerin depolandığı bu yol MapReduce prensibine benzemektedir. MapReduce prensibinde kelime anahtarı ve kelimenin konumu da depolama adresini temsil etmektedir. Bu büyük veride eşleme (mapping) olarak adlandırılmaktadır. Yer değiştirme aşamasında kimlik saptayıcıların tespit edilmesi ve depolanması bittikten sonra orijinal veride bellekteki her bir kelimenin karıştırılması ve yer değiştirilmesi ile kimliksizleştirme adımı başlar. Yer değiştirme aşaması şu şekilde çalışır: bir kez tanınan antikor sahibinin belleğinde depolanır. Bu birçok tekrarlama ile devam eder. Her birinde iki antikor rastgele olarak seçilir ve sıralanır. Sonrasında sıralanan kelimeden iki rastgele bileşen seçilir ve sırası değiştirilir. Ardından sonuç iki bölüm olacak şekilde bir rastgele konum alınır. İlki seçilen ilk antikor ile ikincisi ise ikinci ile yer değiştirilir. Sonuç olarak orijinal metinde yer değiştirme antikorları kimlik saptayıcılarını değiştirmek için çözülür [14].

Tekrardan kimlik saptama kimliksizleştirilmiş belgelerin her bir kelimenin kendi haline yeniden sokulmasıyla orijinal yapısına döndürülmesidir. Bunu yapmanın yolu depolama aşamasında kullanılan matematiksel fonksiyon kadar kolaydır. Kesme işlemi bu aşamada yapılır. Hafızadaki her bir antikor (antibody) her bir kelimenin konumunu belirleyen bir anahtardır. Antikorların konumları belirlendikten sonra sistem; metin dosyalarına eklenecek olan hakiki kelimeleri oluşturmak için bu antikorları çözer (decode) [14].

Büyük verinin gelişimini önleyen en büyük engellerden biri kullanıcıların mahremiyetidir. Bu konu bağlamında birçok gelişmiş araştırma yapılmıştır. Bunlardan biri ise homomorfik şifreleme olarak bilinen kriptografiyle ilgili bir kavramdır. Homomorfik şifreleme, şifre çözme ihtiyacı olmadan şifreli veri üzerinde işlemlerin uygulanmasını sağlar. Genetik algoritmalar son zamanlarda öne sürülen kriptosistemlerinin yeni bir türüdür ve veri madenciliği evrimsel yöntemlerini kullanan değiştirme (substitution) ve yer değiştirme (transposition) gibi klasik kriptografi tekniklerini geliştirmeyi amaçlamaktadır. Bu tip sistemlerin etkinliği IND-CPA (attacks with chosen plaintext) ve IND-CCA (attacks with chosen ciphered text) ile kanıtlanmıştır. Çalışmada TSZ (To, Safavi-Naini ve Zhang) olarak bilinen homomorfik kriptosisteminin etkinliği TSZ ve evrimsel kriptografiyi birleştiren ve iki yöntemin avantajlarını kullanan yeni bir yaklaşım öne sürülerek artırılmıştır [15].

Bulut bilişim; sağlık ve işletme gibi sektörlerdeki çeşitli büyük veri uygulamalarının farklı aşamalarını desteklemek için ölçeklenebilir IT altyapıları sağlamaktadır. Elektronik sağlık kayıtları gibi veri setleri çoğunlukla hassas gizlilik bilgileri içermektedir. Bu bilgilerin bulutta üçüncü şahıslar ile paylaşılması gizlilik endişelerini beraberinde getirir. Veri gizliliği koruması için pratik ve yaygın olarak benimsenen teknik; bir gizlilik modelini sağlayan genelleme (generalization) ile verinin anonimleştirilmesidir. Ancak var olan çoğu gizlilik koruma yaklaşımı yetersizliği veya kötü ölçeklenebilirliği sebebiyle büyük verileri küçük boyutlu veri setlerine dönüştürürler. Zhang ve diğerleri tarafından yapılan çalışmada [16] yakınlık gizlilik ihlallerine karşı büyük veri anonimleştirme için yerel yeniden kodlama problemi (local-recoding problem) incelenmiştir ve bu probleme ölçeklenebilir bir çözüm tanımlamak için çalışılmıştır. Özellikle hassas değerler ve çoklu hassas özelliklerin semantik yakınlığını sağlayan bir yakınlık gizlilik modeli sunulmuştur ve yerel yeniden kodlama problemi bir yakınlık farkındalıklı kümeleme problemi (proximity-aware clustering problem) olarak modellenmiştir. Bu problemleri çözebilmek için t-atalı (ancestors) kümeleme (k-means'e benzer) algoritmasından oluşan ölçeklenebilir iki fazlı kümeleme yaklaşımı ve yakınlık farkındalıklı yığınsal (agglomerative) kümeleme algoritması sunulmuştur. Bulutta veri paralel hesaplama gerçekleştirerek yüksek ölçeklenebilirlik kazanmak için algoritmalar MapReduce ile tasarlanmıştır. Gerçek veri setleri üzerindeki kapsamlı deneyler çalışmadaki yaklaşımın mevcut yaklaşımlar üzerinde yakınlık gizlilik ihlalleri savunma kapasitesini, ölçeklenebilirliği ve yerel yeniden kodlama anonimleştirmesinin zaman verimliliğini artırdığını göstermektedir.

Büyük veri ekonomik büyüme ve teknik yenilik için bilgi madenciliği yapabilmesi sebebiyle son zamanlarda oldukça dikkat çekmektedir ve birçok araştırma çalışması yüksek hacim (volume), velocity (hız) ve çeşitlilik (variety) ("3V" olarak anılır) zorlukları sebebiyle büyük veri işlemeye yönlendirilmiştir. Eğer veri güvenilir (authentic) değilse madencilikten elde edilen yeni bilgi inandırıcı olmayacaktır. Gizlilik iyi bir şekilde sağlanmadıysa insanlar kendi verilerini paylaşmak için isteksiz olabilirler. Çünkü büyük veride güvenlik "gerçeklik (veracity)" olarak adlandırılan yeni bir boyut

olarak incelenmektedir. Lu ve diğerleri tarafından yapılan çalışmada [17] gizlilik açısından büyük verinin yeni zorluklarının kullanılması amaçlanmıştır ve büyük veri çağındaki etkili ve gizlilik korumalı programlama üzerine odaklanılmıştır. İlk olarak büyük veri analitiğinin genel mimarisi biçimlendirilmiştir. Ardından karşılık gelen gizlilik gereksinimleri tanımlanmıştır. Son olarak etkili ve gizlilik koruyucu kosinüs (cosine) benzerlik hesaplama protokolü büyük veri çağında veri madenciliğinin etkinliğine ve gizlilik gereksinimlerine karşılık olarak öne sürülmüştür [17].

Hastane ya da banka gibi bir veri sahibinin gizli olarak saklanan kişiye özgü alanlara ayrılmış veri yığınına sahip olduğunu düşünelim. Bu veri sahibinin verinin bir örneğini araştırmacılarla paylaşmak istediğini varsayalım. Veri sahibi verinin kaynağı olan bireylerin kimliklerinin tekrardan tespit edilemeyecek şekilde bilimsel teminatlı özel verinin örneğini nasıl paylaşmalıdır? Sweeney tarafından yapılan çalışmada [18] öne sürülen çözüm k-anonimlik olarak adlandırılan biçimsel bir koruma modeli ve dağıtım için bir dizi ilkeyi içermektedir. Bir yayın (release) o yayındaki her bir kişinin bilgisi, yine o yayında bilgisi olan en az k-1 birey tarafından ayırt edilemiyorsa k-anonimlik korumasını temin eder. Ayrıca çalışmada k-anonimliğe bağlı kalan yayınlar üzerinde gerçekleştirilebilen tekrardan kimlik saptama atakları incelenmiştir. k-anonimlik koruma modeli gerçek dünya sistemlerine dayandığı için önemlidir.

Birçok kurum halk sağlığı ve nüfus araştırması gibi amaçlarla mikro verilerini yayınlamaktadır. İsim, sosyal güvenlik numarası gibi bireyleri açık bir şekilde belirleyen özellikler genel olarak silinseler de bu veritabanları bazen posta kodu, cinsiyet ve doğum tarihi gibi özellikler üzerinden anonim kalması gereken bireyleri tekrardan belirlemek için diğer açık veritabanları ile birleştirilirler. Birleştirme atakları (joining attacks) Internet üzerinden birbirini tamamlayan diğer veritabanlarının olmasıyla daha kolay gerçekleştirilir [19].

K-anonimleştirme (anonymization) yayınlanan mikro verinin parçalarını genelleyerek ve/veya gizli tutarak birleştirme ataklarını önleyen bir tekniktir. Böylece k boyutundaki bir gruptaki herhangi bir birey eşsiz olarak ayırt edilemez. LeFevre ve diğerleri çalışmasında [19] k-anonimleştirme modelini gerçeklemek için tam alan genelleme (full-domain generalization) olarak adlandırılan uygulanabilir bir çerçeve sunmuştur. Minimum tam alan genellemeyi gerçekleştirmek için bir algoritma dizisi sunulmuştur ve bu algoritmaların iki gerçek veritabanı üzerinde önceki algoritmalardan büyüklük sırasına göre daha hızlı olduğu gösterilmiştir.

Veri kimliksizleştirme araştırma amaçlı verinin yayınlanması ihtiyaçlarını ve bireylerin gizlilik isteklerini uzlaştırmaktadır. Bayardo ve Agrawal çalışmasında [20] k-anonimleştirme olarak bilinen güçlü kimliksizleştirme yöntemi için bir optimizasyon algoritması öne sürmüştür ve değerlendirmelerini yapmıştır. k-anonimleştirilmiş bir veri setinde her bir kayıt en az diğer k-1 kayıttan ayırt edilemez olmalıdır. Optimize edilmiş k-anonimliğin basit kısıtları NP-hard olsa bile önemli hesaplamalı zorluklara neden olmaktadır. Gerçek nüfus sayımı verileri üzerinde yapılan deneylerde öne sürülen algoritmanın iki temsili maliyet ölçümü ve geniş bir k aralığı altında en iyi k-anonimleştirmelerini bulabildiği gösterilmiştir. Ayrıca algoritmanın giriş verisi veya giriş parametrelerinin makul sürede en iyi çözümü bulmayı engellediği durumlarda da iyi anonimleştirme ürettiği gösterilmiştir. Son olarak algoritma; anonimleştirme kalitesi ve performans üzerinde farklı kodlama yaklaşımlarının ve problem varyasyonlarının etkilerini araştırmak için kullanılmıştır. Yazarlara göre bu çalışma problemin genel bir modeli altında çözülmesi zor alan bir veri setinin en iyi k-anonimleştirmesini kanıtlayan ilk sonuçtur.

K-anonimlik mikro veri yayınlamada gizliliği korumak için öne sürülen bir mekanizmadır ve k-anonimliği sağlamak için çok sayıda kodlama modeli düşünülmüştür. LeFevre ve diğerleri tarafından yapılan çalışmada [21] önceki çalışmalarda (tek-boyutlu) görülmeyen ek esneklik derecesi sağlayan yeni bir çok boyutlu (multidimensional) model öne sürülmüştür. Bu esneklik genel amaçlı ölçü birimleri ve daha belirli sorgu cevaplanabilirlik kavramları ile ölçüldüğü gibi daha yüksek kalitede anonimleştirmeye yol açmaktadır. Optimal çok boyutlu anonimleştirme NP-hard'dır. Fakat çalışmada basit bir greedy yaklaşım algoritması öne sürülmüş ve deneysel sonuçlar bu algoritmanın iki tek boyutlu model için ayrıntılı (exhaustive) optimal algoritmadan daha makul anonimleştirme yaptığını göstermektedir.

Sweeney çalışmasında [22] k-anonimliği (anonymity) gerçekleştirmek için genelleme (generalization) ve gizlemeyi (suppression) birleştiren biçimsel bir sunum ortaya koymuştur. Genelleme bir değerin daha az belirli fakat anlamsal olarak tutarlı bir değer ile değiştirilmesi (veya yeniden kodlanması) ile ilgilidir. Gizleme ise bir değerin hiçbir biçimde yayınlanmamasıdır. MinGen (Minimal Generalization Algorithm, Minimum Genelleme Algoritması) teorik bir algoritmadır ve minimum bozulmalı k-anonimlik koruması sağlamak için bu iki tekniği birleştirmektedir.

Gizlilik; sosyal bilim araştırması ya da iş analizi için sosyal ağ verilerinin yayınlanmasında veya paylaşılmasında en önemli sorunlardan biridir. Son zamanlarda araştırmacılar yapı bilgisi yoluyla düğüm kimlik saptamasını engellemek için k-anonimliğe gizlilik modelleri geliştirmişlerdir. Ancak bu gizlilik modelleri uygulansa bile bir düğüm grubu yaygın bir şekilde aynı hassas etiketleri (özellikleri) paylaşıyorsa saldırgan, birinin gizli bilgisine ulaşabilir. Diğer bir değişle burada etiket düğüm ilişkisi saf yapı anonimleştirme yöntemleri ile iyi korunmamaktadır. Ek olarak uç düzenleme ya da düğüm kümelemeye dayanan var olan yaklaşımlar önemli ölçüde anahtar grafik özelliklerini değiştirebilir. Yuan ve diğerleri çalışmasında [23] bireylerin hassas etiketlerinin yanında yapısal bilgi korumasını göz önünde bulunduran bir k-derece-l-çeşitlilik anonimlik modeli (k-degree-l-diversity anonymity model) tanımlamıştır. k-derece-l-

çeşitliliğin gereksinimlerini sağlamak için orijinal grafdan yeni bir graf oluşturmak için gürültü (noise) düğüm ekleme algoritması tasarlanmıştır. Burada orijinal graf üzerinde en az bozulma kısıtlaması vardır. Çalışmada gerçekleştirilen deney sonuçları gürültü düğüm ekleme algoritmasının yalnızca uç düzenleme kullanan çalışmalardan daha iyi sonuçlar verdiğini göstermektedir.

iRODS (Integrated Rule-Oriented Data, Entegre Kural-Odaklı Veri) isimli yeni bir teknik büyük veride güvenliği ve gizliliği sağlamak için öne sürülmüştür [1, 24]. iRODS'un en önemli teknik özellikleri şunlardır: birleşmiş veri sistemleri (grids) veya akıllı bulutlar, dağıtık kural motoru, iCAT metadata kataloğu, ortak erişim sağlayan depolama erişim katmanı, grafiksel kullanıcı arayüzü ve komut-satırı-tabanlı istemcilerin zengin kombinasyonu ve iRODS veri sistemi ile etkileşimi sağlayan API'ler. iRODS birçok veri yönetimi uygulamasında kullanılmıştır ve dünya çapında çok sayıda kuruluş tarafından benimsenmiştir. iRODS teknolojisinin kural tabanlı ve büyük ölçekli veri yönetimindeki çeşitli zorlukların üstesinden gelmek için nasıl uygulandığıyla alakalı birçok yayın vardır [1, 25-28]. iRODS teknolojisi veriyi korumak ve gizliliği garanti etmek için olan yaygın yaklaşımlar üzerinde gelişme sağlamaktadır. Bu gelişmeler şunlardır: güvenlik kontrollerinin kapsamlı seti, veri erişiminin gelişmiş kontrolü ve metadata üzerinden kullanımı, depolama sanallaştırması ve veri güvenliği yaşam döngüsü ve kalıcı tanımlayıcılar (persistent identifiers) [1]. Tablo 1'de büyük veri güvenliği için geliştirilen yöntemler özetlenmektedir.

Tablo 1. Büyük veri güvenliği için geliştirilen yöntemler (atıf sayısına ve kronolojiye göre)
(The methods developed for big data security (in terms of the number of citation and chronology))

Yayın İsmi	Yazar	Yıl	Yöntem	Atıf Sayısı
k-anonymity: a model for protecting privacy	Sweeney [18]	2002	*k*-anonimlik olarak adlandırılan biçimsel bir koruma modeli öne sürülmüştür.	4179
Achievıng k-anonymity privacy protection using generalization and suppression	Sweeney [22]	2002	*k*-anonimliği gerçekleştirmek için genelleme (generalization) ve gizlemeyi (suppression) birleştiren biçimsel bir sunum ortaya konulmuştur. MinGen (Minimal Generalization Algorithm) teorik bir algoritmadır ve minimum bozulmalı *k*-anonimlik koruması sağlamak için bu iki tekniği birleştirmektedir.	1601
Incognito: Efficient Full-Domain K-Anonymity	LeFevre ve diğ. [19]	2005	*k*-anonimleştirme modelini gerçeklemek için tam alan genelleme olarak adlandırılan uygulanabilir bir çerçeve sunulmuştur ve iki gerçek veritabanı üzerinde önceki algoritmalardan hızlı olduğu gösterilmiştir.	1168
Data privacy through optimal k-anonymization	Bayardo ve Agrawal [20]	2005	*k*-anonimleştirme olarak bilinen güçlü kimliksizleştirme yöntemi için bir optimizasyon algoritması öne sürülmüş ve değerlendirilmiştir. Bu çalışma problemin genel bir modeli altında çözülmesi zor alan bir veri setinin en iyi *k*-anonimleştirmesini kanıtlayan ilk sonuçtur.	1125
Mondrian Multidimensional K-Anonymity	LeFevre ve diğ. [21]	2006	Önceki çalışmalarda (tek-boyutlu) görülmeyen ek esneklik derecesi sağlayan yeni bir çok boyutlu (multidimensional) model öne sürülmüştür.	1003
Toward efficient and privacy-preserving	Lu ve diğ. [17]	2014	Gizlilik açısından büyük verinin yeni zorluklarının kullanılması amaçlanmıştır ve büyük veri çağındaki etkili ve	86

computing in big data era			gizlilik korumalı programlama üzerine odaklanılmıştır.	
Protecting Sensitive Labels in Social Network Data Anonymization	Yuan ve diğ. [23]	2013	Bireylerin hassas etiketlerinin yanında yapısal bilgi korumasını göz önünde bulunduran bir *k*-derece-*l*-çeşitlilik anonimlik modeli tanımlanmıştır.	55
Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud	Zhang ve diğ. [16]	2015	Yakınlık gizlilik ihlallerine karşı büyük veri anonimleştirme için yerel yeniden kodlama problemi (local-recoding problem) incelenmiştir ve bu probleme ölçeklenebilir bir çözüm tanımlamak için çalışılmıştır.	22
Big Data security and privacy: A review	Matturdi ve diğ. [1]	2015	Büyük veri güvenliği ve mahremiyeti hakkında genel bilgi verilmektedir.	18
A Multilayer Evolutionary Homomorphic Encryption Approach for Privacy Preserving over Big Data	Rahmani ve diğ. [15]	2014	TSZ (To, Safavi-Naini ve Zhang) olarak bilinen homomorfik kriptosisteminin etkinliği TSZ ve evrimsel kriptografiyi birleştiren ve iki yöntemin avantajlarını kullanan yeni bir yaklaşım öne sürülerek artırılmıştır.	7
An approach towards big data — A review	Gupta ve Tyagi [11]	2015	Büyük verinin işlevsel güvenliği, siber güvenlikte büyük veri analizleri ve tehdit tespiti için güvenlik analizleri gibi konularda genel bilgi verilmektedir.	5
Big Data: Big Promises for Information Security	Alguliyev ve Imamverdiyev [9]	2014	Büyük veri güvenliği ve siber güvenlik için olan büyük veri zorlukları hakkında genel bilgi verilmektedir.	5
De-identification of Textual Data Using Immune System for Privacy Preserving in Big Data	Rahmani ve diğ. [14]	2015	Özel bir bağışıklık sistemi kullanarak metin verilerini kimliksizleştiren yeni bir model ortaya konulmuştur. Bu yaklaşımda kimliksizleştirme ve tekrardan kimlik saptama olmak üzere iki ana aşama vardır.	2

Bu çalışmada büyük veri güvenliği için geliştirilen yöntemler kısaca açıklanmıştır ve bu alanda çalışacak olan yeni araştırmacılara yol göstermek hedeflenmiştir. İncelenen makalelere bakılarak bir değerlendirme yapılmak istenildiğinde büyük veri üzerindeki güvenlik ve kişi mahremiyeti mekanizmalarının henüz olgunlaşmamış olduğu ve sürekli olarak geliştiği görülmektedir.

6. SONUÇ (CONCLUSION)

Bu çalışmada büyük veride kişi mahremiyetinin korunması ve büyük veri güvenliği ile ilgili bugüne kadar yapılmış olan bilimsel çalışmalar özetlenmiştir. Ayrıca büyük veri sistemlerini korumak için gerekli olan araçlardan ve bu sistemlerin güvenliğinin sağlanmasında karşılaşılan zorluklardan bahsedilmiştir. Buna ek olarak büyük veri kaynaklarının neler olduğu açıklanmıştır.

Günümüzde medya paylaşımları, sosyal ağ gibi farklı kaynaklardan günlük olarak birçok veri üretilmektedir ve veri miktarı hızlı bir şekilde artmaktadır. Bu da büyük veri kavramını ortaya çıkarmıştır ve kişi mahremiyeti noktasında farklı birçok zorluğa sebep olmuştur. Bu alanda yapılan çalışmaların giderek artması gelecekte bu konunun daha da önem kazanacağını göstermektedir.

TEŞEKKÜR (ACKNOWLEDGEMENT)

Bu çalışma İstanbul Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'na bağlı olarak yürütülen "Büyük Veride Etkin Gizlilik Koruması için

Yazılım Tasarımı" adlı doktora tezinin bir bölümüdür. Ayrıca İstanbul Ticaret Üniversitesi Bilişim Teknolojileri Uygulama ve Araştırma Merkezi tarafından desteklenmektedir.

KAYNAKLAR (REFERENCES)

[1] B. Matturdi, X. Zhou, S. Li, F. Lin, "Big Data security and privacy: A review", *China Communications*, 11(14), 135-145, April 2015.

[2] N. Perlroth, J. Larson, S. Shane, **NSA able to foil basic safeguards of privacy on web**, The New York Times, 6, 5 September 2013.

[3] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization", *UCLA Law Review*, 57, 1701, 2010.

[4] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, "Mondrian multidimensional k-anonymity", **Proceedings of the 22nd International Conference on Data Engineering**, IEEE, 25-25, 03-07 April 2006.

[5] L. Lü, T. Zhou, "Link prediction in weighted networks: The role of weak ties", *EPL (Europhysics Letters)*, 89(1), 18001, January 2010.

[6] A. Clauset, C. Moore, M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks", *Nature*, 453, 98-101, May 2008.

[7] D. Yin, L. Hong, X. Xiong, "Link formation analysis in microblogs", **Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval**, 1235-1236, 2011.

[8] R. N. Lichtenwalter, J. T. Lussier, N. V. Chawla, "New perspectives and methods in link prediction", **Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining**, 243-252, 2010.

[9] R. Alguliyev, Y. Imamverdiyev, "Big Data: Big Promises for Information Security", **2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)**, 1-4, Astana, 15-17 October 2014.

[10] **Cloud Security Alliance (CSA): Expanded Top Ten Big Data Security and Privacy Challenges**, April 2013.

[11] P. Gupta, N. Tyagi, "An approach towards big data — A review", **2015 International Conference on Computing, Communication & Automation (ICCCA)**, 118-123, Noida, 15-16 May 2015.

[12] Internet: Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments, https://securosis.com/assets/library/reports/SecuringBigData_FINAL.pdf, 22.09.2016.

[13] T. Mahmood, U. Afzal, "Security Analytics: Big Data Analytics for Cybersecurity", **2013 2nd National Conference on Information Assurance (NCIA)**, 2013.

[14] A. Rahmani, A. Amine, M. R. Hamou, "De-identification of Textual Data Using Immune System for Privacy Preserving in Big Data", **2015 IEEE International Conference on Computational Intelligence & Communication Technology (CICT)**, 112-116, Ghaziabad, 13-14 February 2015.

[15] A. Rahmani, A. Amine, R. H. Mohamed, "A Multilayer Evolutionary Homomorphic Encryption Approach for Privacy Preserving over Big Data", **2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)**, 19-26, Shanghai, 13-15 October 2014.

[16] X. Zhang, W. Dou, J. Pei, S. Nepal, C. Yang, C. Liu, J. Chen, "Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud", *IEEE Transactions on Computers*, 64(8), 2293-2307, 2015.

[17] R. Lu, H. Zhu, X. Liu, J. K. Liu, J. Shao, "Toward efficient and privacy-preserving computing in big data era", *IEEE Network*, 28(4), 46-50, 2014.

[18] L. Sweeney, "k-anonymity: a model for protecting privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 557-570, 2002.

[19] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, "Incognito: Efficient Full-Domain K-Anonymity", **ACM SIGMOD international conference on Management of data**, 49-60, Baltimore, Maryland, USA, 14-16 June 2005.

[20] R. J. Bayardo, R. Agrawal, "Data privacy through optimal k-anonymization", **21st International Conference on Data Engineering**, IEEE, 217-228, 5-8 April 2005.

[21] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity", **22nd International Conference on Data Engineering**, IEEE, 25, 03-07 April 2006.

[22] L. Sweeney, "Achievıng k-anonymity privacy protection using generalization and suppression", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 571-588, 2002.

[23] M. Yuan, L. Chen, P. S. Yu, T. Yu, "Protecting Sensitive Labels in Social Network Data Anonymization", *IEEE Transactions on Knowledge and Data Engineering*, 25(3), 633-647, 2013.

[24] M. Jensen, "Challenges of privacy protection in big data analytics", **2013 IEEE International Congress on Big Data (BigData Congress)**, 235-238, 2013.

[25] A. Rajasekar, R. Moore, C. Hou, "iRODS Primer: integrated rule-oriented data system", *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1), 1-143, 2010.

[26] A. Rajasekar, R. Moore, M. Wan, "Applying rules as policies for large-scale data sharing", **2010 International Conference on Intelligent Systems, Modelling and Simulation (ISMS)**, IEEE, 322-327, 2010.

[27] I. Barg, D. Scott, E. Timmermann, "NOAO E2E integrated data cache initiative using iRODS", **XX. ASP Conference Proceedings Astronomical Data Analysis Software and Systems**, 442, 497-500, 2011.

[28] J. L. Schnase, W. P. Webster, L. A. Parnell, "The NASA Center for Climate Simulation Data Management System", **2011 IEEE 27th Symposium on Mass Storage Systems and Technologies (MSST)**, 1-6, 2011.