Veri Akış Uzunluğu Kestiriminde Örnekleme Stratejilerinin Karşılaştırılması
Comparison of Sampling Strategies for Flow Length Estimation

M. Özgün Demir[1], Barış Kurt[2], Saliha Büyükçorak[1], Güneş Karabulut Kurt[1], A. Taylan Cemgil[2], and Engin Zeydan[3]

[1]İstanbul Teknik Üniversitesi, {demirmehmet1235, buyukcorak, gkurt}@itu.edu.tr
[2]Boğaziçi Üniversitesi, {baris.kurt, taylan.cemgil}@boun.edu.tr
[3]AveaLabs, Cakmak Mah. Balkan Cad. No: 49, Umraniye, Istanbul, {Engin.Zeydan}@avea.com.tr

***Özetçe* —Ağ trafiğinin analizi ve anlamlandırılması; kullanıcı davranışlarının belirlenmesi, kaynak kullanımı gibi konularda sağladığı öngörüler nedeniyle oldukça önemlidir. Söz konusu parametreleri belirlemek üzere gerçekleştirilen veri analizlerini; veri gizliliği, ağır işlem/hafıza yükü gibi sıkıntıları beraberinde getiren ağdan akan tüm verileri kullanarak yapmak yerine, akış verisinin küçük bir alt kümesi seçilerek, bir başka deyişle örnekleme yapılarak tüm veri hakkında kestirimler yapılarak gerçekleştirilmek istenir. Bu çalışmada, paket ve zaman bazlı eş aralıklı örnekleme, paket ve zaman bazlı rastgele örnekleme olmak üzere dört farklı örnekleme stratejisi incelenerek paket akış uzunluk dağılımları kestirilmiş ve gerçek uzunluk dağılımı ile karşılaştırılmıştır. Analiz edilen veri üzerinde stratejiler arasında büyük farklılıklar gözlemlenmemiştir.**

Anahtar Kelimeler*—*Veri akışı, eş aralıklı ve rastgele örnekleme, akış uzunluk dağılımı kestirimi

***Abstract*—The importance of the analysis and understanding of the network traffic has constantly been increasing due to insights that this provides towards determination of user behaviour and resource usage. The data analyses in order to determine the related parameters are performed by selection of a small subset of the complete flow data due to data privacy and heavy computational/memory load issues. That is, sampling is required in order to detect the properties of the complete data set. In this work, four distinct sampling schemes, namely the packet based uniform sampling, time-slot based uniform sampling, packet based random sampling and time-slot based random sampling are investigated from which packet flow length distributions are estimated and compared with the actual data. No major differences are observed amongst the strategies based on the analysed data.**

Keywords*—*Data flow, uniform and random sampling, flow size distribution estimation

I. GİRİŞ

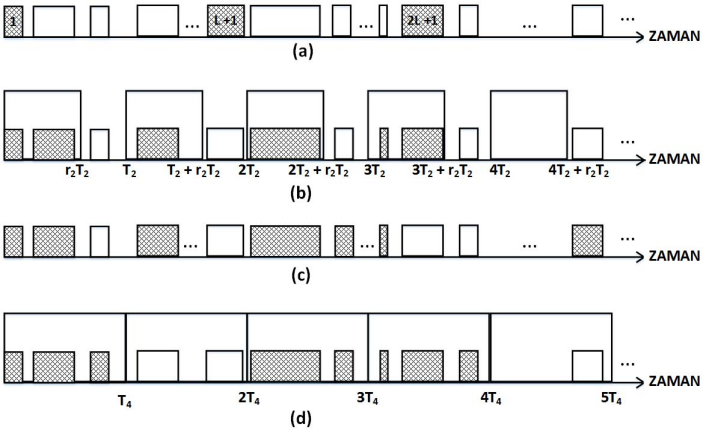
Günümüzde, gelişen bilgisayar ve akıllı telefon teknolojileri ile bu sistemleri kullanan kullanıcı sayısı ve internet kullanımına olan talep çarpıcı bir şekilde artmıştır. Yükselen bu talep doğrultusunda bilgi aktarımı arttığından büyük miktarlarda verilerin kullanıcılara iletilmesi gereksinimi doğmuş ve bunun sonucunda haberleşme sistemlerinde hafıza ve işlem yükü problemleri ortaya çıkmıştır.

İnternet üzerinde bilgi aktarımı, paket tabanlı iletişim ile sağlanır. İki nokta arasında akan paket bütününe ise akış (flow) adı verilir. Bu çalışmada [1]'de verilen flow tanımı esas alınmış olup, buna göre akış kavramı, kaynak ve hedef IP adresleri, kapı (port) numaraları, kullanılan internet protokolü (IP) olmak üzere belirtilen 5 değere göre sınıflandırılır. Tanımlanan bu akış bilgilerine dayanarak yapılan trafik analizleri, servis sağlayıcıya kullanıcı davranışları, kaynak kullanımı, trafik yoğunluğu gibi konularda detaylı bilgiler vererek gerekli ağ operasyonlarının yapılmasında önemli görev üstlenir [2].

Sistemlerin bahsedilen hafıza ve işlem yükü problemlerinden etkilenmemesi için iyi tasarımlar ve akılcı çözümler gerekmektedir. Örnekleme, bu alanda uygulanan en önemli yöntemlerden biridir [3]. Bu doğrultuda, bu çalışma kapsamında çeşitli örnekleme tekniklerinin veri akış boyutu kestirimine olan etkisi incelenerek yöntemler karşılaştırılmıştır.

Veri trafiğinin büyük hacimli ve yüksek hızda varış oranlarına sahip olmasının getirdiği problemlere bir çözüm olarak paket örnekleme modelleri sunulmuş ve böylece tamamlanmamış verilerden veri akış büyüklüğü dağılımlarının tahmin edilmesi hedeflenmiştir [4]–[6]. [7]'da ise, ilgili yazında yer alan örnekleme tiplerinden Poisson ve düzgün örnekleme tipleri karşılaştırılmıştır. Kumar ve diğerleri de, örneklenmiş veriden akış dağılımlarını kestirmeyi amaçlayan çalışmalar yapmış ve tüm veriyi örneklenmiş paketler ile birleştirerek, akış dağılımlarını yüksek başarım ile bulmuşlardır [8]. Grieco ise yaptığı çalışmada akış dağılımını kestirmenin yerine gerçek zamanlı bir algoritma tasarlayarak örneklenmiş paket trafiğinin spektral özelliklerini incelemiştir [9]. Bu çalışma kapsamında ise, zaman ve paket bazında düzgün ve rastgele örnekleme tiplerinin veri akış uzunluğu kestirimi üzerine etkisi incelenmiş ve buna ek olarak akış varışlarının microsaniye çözünürlüğünde Poisson sürece uygunluğu dalgacık katsayıları kullanılarak test edilmiştir.

978-1-4673-5563-6/13/$31.00



Şekil 1: Örnekleme teknikleri. (a) Paket bazlı eş aralıklı örnekleme (b) Zaman bazlı eş aralıklı örnekleme (c) Paket bazlı rastgele örnekleme (d) Zaman bazlı rastgele örnekleme (taralı kutucuklar örneklenen paketleri, boş olan kutular atlanan paketleri göstermektedir)

II. ÖRNEKLEME TEKNİKLERİ

Örnekleme işleminde ağ trafiğindeki paketlerin bazıları belirli bir stratejiyle seçilir. Seçilen paketlerin uzunlukları ait olduğu akışın toplam uzunluğuna eklenir ve her akışın bir parçası gözlemlenmiş olur. Akışın gerçek uzunluğu ise örnekleme sonrasında örnekleme yöntemine bağlı bir kestirim ile belirlenir. Bu çalışmada, aşağıda açıklanan dört farklı örnekleme yöntemi kullanılmıştır.

1) Paket Bazında Eş Aralıklı Örnekleme: Paket bazında eş aralıklı örnekleme sabit bir L periyodu kullanarak ağ üzerinde taşınan her L paketten birini gözlememizle elde edilir (Şekil 1a). Bu örneklemede bir paket gözlemledikten sonra sıradaki $L-1$ paket atlanır. Dolayısıyla her akışın uzunluğu eksik hesaplanır, hatta az paket sayılı akışların tüm paketleri atlanabilir ve dolayısıyla sayılmazlar.

2) Zaman Bazında Eş Aralıklı Örnekleme: Zaman bazında eş aralıklı örneklemede ise sabit bir T periyoduyla, zaman içerisindeki her T sürede bir, n tamsayı olmak üzere $[nT, p_2 nT]$ zaman aralığındaki tüm paketler gözlemlenerek gerçekleşir (Şekil 1b).

3) Paket Bazında Rastgele Örnekleme: Paket bazında rastgele örneklemede her bir paket p_3 olasığıyla gözlemlenir, ya da $(1-p_3)$ olasılığıyla göz ardı edilir (Şekil 1c).

4) Zaman Bazında Rastgele Örnekleme: Zaman bazında rastgele örneklemede, zaman ekseni eşit aralıklara bölünüp, bu aralık içinde kalan paketler p_4 olasılığıyla gözlemlenir (Şekil 1d).

III. AKIŞ UZUNLUĞU DAĞILIMI KESTIRIMI

Bu çalışmamda örneklenen veriden gerçek akış uzunluğu dağılımını kestirmek için verinin en büyük olabilirlik kestirimi uyguladık. Kestirmek istediğimiz bu dağılıma $\phi_{1:K}$, örnekleme süresince oluşan akış setine ise $x_{1:N}$ diyelim. Burada N akışların sayısı, K ise en uzun akışın boyudur. Bu durumda herhangi bir akışın k adet paketten oluşma ihtimali $P(x_n = k) = \phi_k$ olacaktır.

Örnekleme sonucunda bazı akışlardan hiç paket örnekleyemeyeceğimiz için, R adet akış gözlemlemiş olacağız. Bu gözlem setine $y_{1:R}$ diyelim. Amacımız y gözlemlerinden ϕ dağılımını kestirmek olduğundan $p(y|\phi)$ değerini ϕ'ye göre en büyütmemiz gerekir. Bu olasılık gözlemlenemeyen $x_{1:N}$ değerlerine bağlıdır.

$$\log p(y|\phi) = \log \sum_x p(y|x)p(x|\phi)$$

Bu ifadeyi enbüyütmek için beklenti en büyütme (BE) algoritması kullandık [2].

$$Q(\phi, \phi^{old}) = \langle \log p(x, y|\phi) \rangle_{p(x|y,\phi^{old})}$$

B adımı:

$$p(x_n = i|y_n, \phi^{old}) = \frac{p(y_n|x_n = i)\phi_i^{old}}{\sum_k p(y_n|x_n = k)\phi_k^{old}}$$

E adımı:

$$\phi_k = \frac{\sum_{n=1}^{N} p(x_n = k|y_n, \phi^{old})}{\sum_{n=1}^{N}\sum_{l=1}^{K} p(x_n = l|y_n, \phi^{old})}$$

Burada dikkat edilecek olursa, gerçek akışların sayısı, N de gözlemlenememektedir, dolayısıyla bir rastgele değişkendir. Akışların, parametresini (λ) daha önce kestirdiğimiz bir Poisson sürecinden geldiğini varsaydık ve algoritmanın başında N değerini bu dağılıma göre en olası değer olarak sabitledik.

Bir diğer önemli nokta ise, bu makalede $p(y_n|x_n = i)$ dağılımı için tüm örnekleme modellerinde π olasılıklı ikiterimli dağılımı kullandık. Bu dağılım paket bazlı örneklemelerde tam doğru olsa da, zaman bazlı örneklemeler için tam doğru değildir ancak makul sonuçlar ortaya çıkarmıştır.

A. Performans Ölçümü

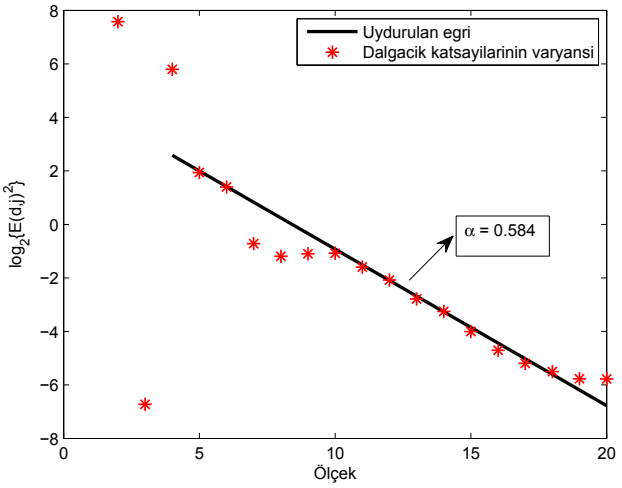
Bu çalışmada gerçek ve kestirilen akış uzunluk dağılımları arasındaki farklılık, hata vektörü tabanlı tekniklerden sıklıkla kullanılan ağırlıklı ortalama bağıl fark (WMRD) metriği kullanılarak ölçülmüştür. WMRD, veri setinin boyutundan bağımsız olarak, iki dağılım arasındaki uzaklığın ölçüsünü verir ve akış uzunluk dağılımları için şu şekilde hesaplanır:

$$\text{WMRD} = \frac{\sum_i |\theta_i - \theta_i^*|}{\sum_i (\theta_i + \theta_i^*)/2}. \quad (1)$$

WMRD değeri için kesin bir standart olmamasına rağmen, bu değerin 1'den küçük olması tercih edilmektedir [10].

B. Geliş Süreci

Bir servisi kulanan kullanıcıların servis sağlayıcıdan o servisi talep etmesi olarak görülebilen geliş süreci haberleşme sistemleri tarafından sağlanan ses ve veri iletimi servislerinin her ikisinde de Poisson süreci ile modellenmektedir. Belirli bir zaman aralığında gelen akışlar ile çeşitli örnekleme senaryoları üzerinde çalıştığımız bu çalışma kapsamında da [11] çalışmasında olduğu gibi akışların geliş süreci Poisson varsayılmıştır. Bu varsayımın gerçekliğini test



Şekil 2: Dalgacık analiz sonuçları

edebilmek amacıyla dalgacık tabanlı istatistiksel bağımsızlık testleri yapılmıştır.

$x(t)$, akış geliş zamanlarını belirten işaret olmak üzere, $x(t)$ çoklu-çözünürlük analizine tabi tutularak zaman-ölçek düzleminde örneklendikten sonra ayrık zamanlı dalgacık dönüşümü alınarak dalgacık katsayıları

$$d(j,k) = \int_{-\infty}^{\infty} x(t) 2^{-j/2} \psi(2^{-j}t - k)dt = \int_{-\infty}^{\infty} x(t)\psi_{j,k}(t)dt \quad (2)$$

kullanılarak hesaplanır [12]. Burada, j oktavı, k çevirici (translation), $\psi(\cdot)$ ana dalgacığı, $\psi_{j,k}(\cdot)$ dalgacık olarak isimlendirilen dönüşümün temel fonksiyonunu ifade etmektedir. Yukarıdaki eşitlik kullanılarak hesaplanan katsayıların varyansının logaritmik değerleri

$$\log_2\left(E\{d(j,k)^2\}\right) = \log_2\left(\frac{1}{l_j}\sum_{k=1}^{l_j} d(j,k)^2\right) = \alpha j + c \quad (3)$$

şeklinde tanımlanır ve oktav j ile arasındaki ilişkinin doğrusal eğimi olarak isimlendirilen α kullanılarak söz konusu bağımsızlık testi gerçekleştirilir [12]. Burada, l_j, j'inci oktavdaki dalgacık katsayılarının sayısını ifade etmektedir.

IV. DENEYSEL SONUÇLAR

A. Veri Detayları

Çeşitli örnekleme senaryolarının veri akış boyutu kestirimi üzerine etkisini incelemek üzere 468938 paket ve 37165 akıştan oluşan bir veri seti üzerinde üzerinde çalışılmıştır.

B. Geliş Süreci Analizi

Akışların geliş sürecinin Poisson sürecine uygunluğunu test edebilmek için gerçekleştirdiğimiz analizlerde, öncelikle akışlardan oluşan bilgi dizisinin ayrık dalgacık transformu alınarak dalgacık katsayıları ve bu katsayıların varyanslarının logaritmik değerleri hesaplanmıştır. Sonrasında, doğrusal eğim olan α değerini bulabilmek için lineer bir eğri uydurulmuş ve elde edilen sonuçlar Şekil 2'de verilmiştir. Microsaniye zaman çözünürlüğünde gerçekleştirilen analiz sonucunda, sıfır eğimli doğrusal bir eğri elde edilemediği söz konusu şekilde de görülmekte olup akışların gelişlerinin birbirinden bağımsız olmadığı ve dolayısıyla akış trafiğinin geliş sürecinin Poisson sürecine uymadığı gözlenmiştir. Bu sonuca rağmen, [11]'de olduğu gibi gelişlerin Poisson sürece uyduğu varsayılmış ve bu sürece ait geliş λ değerleri Tablo I'de verilmiştir.

Tablo I: Kestirilen varış oranı değerleri

Gözlem Süresi	$\hat{\lambda}_1$	$\hat{\lambda}_2$
334838611 ns	1.1099 x 10^{-4}	1.1099 x 10^{-4}

C. Örnekleme Aşamaları ve Sonuçları

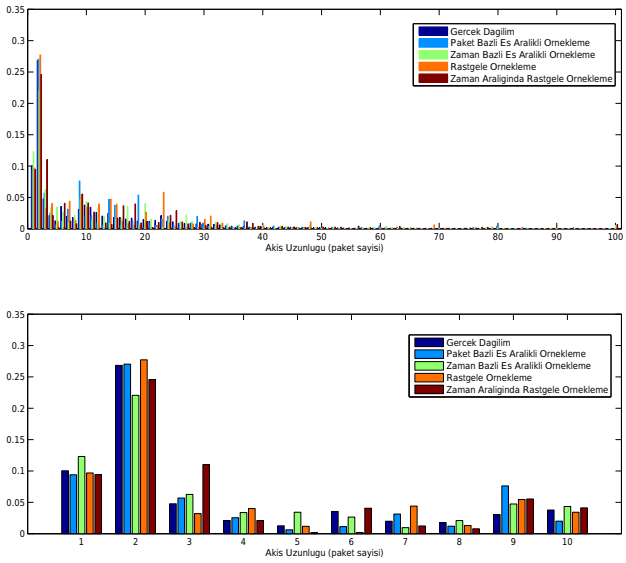
Eş aralıklı örnekleme temel anlamda, örnekleme işleminin herhangi bir düzen altında yapılmasına dayanır. Bu bağlamda, yapılan eş aralıklı örnekleme çalışmaları iki farklı teknikle gerçekleştirilmiştir. Bunlardan ilki, paket bazında eş aralıklı örnekleme diğeri ise zaman bazında eş aralıklı örneklemedir. Rastgele örnekleme ise, zaman ve paket bazlı olmak üzere ikiye ayrılır. Bu örnekleme teknikleri daha ayrıntılı olarak sırasıyla açıklanacaktır.

1) Paket Bazında Eş Aralıklı Örnekleme : Paket bazında eş aralıklı örnekleme esnasında, paketler geliş sıraları bozulmadan, sabit bir örnekleme periyoduna göre örneklenmiştir. Yapılan bu çalışmada örnekleme periyotları 2, 4, 8, 16, 32 ve 64 örnek olacak şekilde belirlenmiştir. Örnekleme işlemi, başlangıç noktası her aşamada teker teker olmak üzere, toplamda örnekleme periyodu kadar kaydırılarak tekrar edilmiştir. Örneklenmiş paketlerden elde edilen, akışların sahip oldukları paket sayıları ve uzunluklarına göre dağılımları Şekil 3'te gösterilmektedir. Bir sonraki aşamada, elde ettiğimiz örneklere ait WMRD değerleri bulunmuş ve elde edilen bu WMRD değerlerinin, ilgili örnekleme periyotlarına göre ortalaması alınmıştır. Bu işlemler sonucunda ulaşılmış değerler Tablo II'de verilmiştir.

2) Zaman Bazında Eş Aralıklı Örnekleme: Zaman bazında yapılan eş aralıklı örnekleme metodunda ise, paketlerin birbirleri arasındaki zaman farkı korunup, ardından örnekleme periyodu olan $T = 2$ ms'lik dilimler içerisinde, p_2T genişlilikli, τ_2 kadar ötelenmiş örnekleme zaman aralıkları seçilmiştir ve bu alanda kalan paketler örneklenmiştir. Burada p değeri sabit bir değişken olup, p_2 için kullanılacak değer sırasıyla paket bazlı yöntemde örnekleme periyodu olan $2, 4, 8, 16, 32, 64$ değerlerinin çarpmaya göre tersi olacak şekilde seçilmiştir. τ_2 ise bir rastlantı değişkenidir. Dolayısıyla bu örneklemede, belirli bir kural çerçevesinde,

Tablo II: Kullanılan örnekleme tekniklerinin WMRD hata değerleri

Örnekleme Oranı	Paket Bazlı Eş Aralıklı Örnekleme	Zaman Bazlı Eş Aralıklı Örnekleme	Paket Bazlı Rastgele Örnekleme	Zaman Bazlı Rastgele Örnekleme
1/2	0.4227	0.5754	0.4600	0.4594
1/4	0.5135	0.8588	0.6370	0.7722
1/8	0.7435	0.8879	0.6237	0.9157
1/16	0.8400	0.8692	0.8190	0.9803
1/32	0.8174	0.9945	0.7856	1.0234
1/64	0.8393	1.1787	0.7978	1.1437



Şekil 3: Gerçek ve kestirilen akış uzunluğu dağılımları. Üstte $\phi_{1:100}$, altta ise $\phi_{1:10}$ gösterilmiştir. Kestirimlerde eş aralıklı örneklemeler için periyodlar $L = 2, T = 2$, rastgele örnekleme için olasılık değeri $p = 0.5$ olacak şekilde seçildi.

rastlantısal özellikler de gözükmektedir. Olasılık işlevin varlığından dolayı, daha geçerli sonuçlar elde etmek için yapılan örnekleme işlemleri 100 defa tekrarlanmıştır. Zaman bazlı eş aralıklı örnekleme sonucu, gözlenen akışların paket sayıların ve uzunluklarına göre dağılımı Şekil 3'te verilmiştir. Paket bazlı yöntemde olduğu gibi, zaman bazlı eş aralıklı örneklemede de, örneklenmiş ile tüm veri dağılımları arasındaki fark WMRD değeri bulunarak belirtilmiştir. Bu aşamada elde edilen WMRD değerlerinin ortalama değerleri Tablo II'de verilmiştir.

3) Paket Bazında Rastgele Örnekleme: Eş aralıklı örneklemenin aksine, paket bazında rastgele örnekleme yönteminde örneklenen paketler, geliş sıralarına göre gözlenip p_3 olasılığına göre rastgele seçilmektedir. Örnekleme işleminin ardından, diğer örnekleme tiplerinde olduğu gibi WMRD değerleri incelenmiştir. Tüm bu işlemler bir rastlantısal sürecin parçası olduğu için örnekleme işlemi her bir p_3 değeri için tüm veri seti boyunca 100 defa tekrarlanmıştır. Yapılan bu işlem, sonuçların genelleşmesini sağlamıştır. Örnekleme sonucu akışların paket sayılarına ve uzunluklarına göre dağılımları Şekil 3'te verilmiştir ayrıca bulunan WMRD değerleri de Tablo II'de verilmiştir.

4) Zaman Bazında Rastgele Örnekleme: Bu örnekleme tipinde ise, zaman ekseni eş aralıklara bölünmüş ve her bir aralık içinde kalan paketler p_4 olasılığına göre örneklenmiştir. Burada p_4 değeri paket bazlı rastgele örneklemeyle aynıdır ve benzer şekilde 100 defa tekrar etmektedir. Zaman bazlı eş zamanlı örneklemeye benzerliği ise örnekleme aralıklarının rastlantısal τ_4 değeri kadar ötelenerek seçilmesidir. Aralarındaki temel fark, zaman bazında rastgele örnekleme yönteminde, aralıkların aralıksız olarak seçilip bu aralıktaki paketlerin belli bir olasılığa göre seçilmesidir. Diğer örnekleme tiplerinde olduğu gibi, bu örnekleme yöntemine ait akış dağılımları Şekil 3'te gösterilmiş, ilgili WMRD değerleri Tablo II'de verilmiştir.

V. SONUÇLAR VE VARGILAR

Bu çalışmada farklı örnekleme teknikleri kullanarak ağ akış uzunluğu dağılımı kestirmeye çalışılmıştır. Tablo II'de verilen WMRD değerlerinden görüldüğü üzere, yüksek oranda örnekleme yapıldığında bütün örneklemeler iyi sonuçlar vermiştir. Örnekleme oranı düştüğünde ise zaman bazlı örneklemeler daha kötüye gitmiştir.

İleriki çalışmalarımızda zaman bazlı örneklemeler için beklenti enbüyütme algoritmasını güncellemeyi ayrıca gözlemleyemediğimiz gerçek veri uzunluğunu Bayesçi yaklaşımla kestirmeyi planlıyoruz.

KAYNAKÇA

[1] A. Kobayashi, B. Claise, H. Nishida, C. Sommer, and F. Dressler, "IP Flow Information Export (IPFIX) Mediation: Problem Statement," IETF, RFC 5982, August 2010.

[2] L. Yang and G. Michailidis, "Sampled based estimation of network traffic flow characteristics," in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, May 2007, pp. 1775–1783.

[3] G. Cheng, J. Gong, and Y. Tang, "A hybrid sampling approach for network flow monitoring," in *End-to-End Monitoring Techniques and Services, 2007. E2EMON '07. Workshop on*, Yearly 2007, pp. 1–7.

[4] N. Duffield, C. Lund, and M. Thorup, "Learn more, sample less: control of volume and variance in network measurement," *Information Theory, IEEE Transactions on*, vol. 51, no. 5, pp. 1756–1775, May 2005.

[5] N. Duffield, "Fair sampling across network flow measurements," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 1, pp. 367–378, Jun. 2012.

[6] B. F. Ribeiro, D. F. Towsley, T. Ye, and J. Bolot, "Fisher information of sampled packets: an application to flow size estimation," in *Internet Measurement Conference*, 2006, pp. 15–26.

[7] M. Roughan, "A comparison of poisson and uniform sampling for active measurements," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 12, pp. 2299–2312, Dec 2006.

[8] A. Kumar, M. Sung, J. J. Xu, and E. W. Zegura, "A data streaming algorithm for estimating subpopulation flow size distribution," *SIGMETRICS Perform. Eval. Rev.*, vol. 33, no. 1, pp. 61–72, Jun. 2005.

[9] L. A. Grieco and C. Barakat, "An analysis of packet sampling in the frequency domain," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '09. New York, NY, USA: ACM, 2009, pp. 170–176.

[10] A. Kumar, M. Sung, J. Xu, and J. Wang, "Data streaming algorithms for efficient and accurate estimation of flow size distribution," in *Joint Int. Conf. on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS/Performance, 2004.

[11] N. Hohn and D. Veitch, "Inverting sampled traffic," *Networking, IEEE/ACM Transactions on*, vol. 14, no. 1, pp. 68–80, Feb 2006.

[12] N. Cackov, "Wavelet-based estimation of long-range dependence in video and network traffic traces," Ph.D. dissertation, Simon Fraser University, 2005.