Машин сургалт ашиглан Tor сүлжээг ангилах

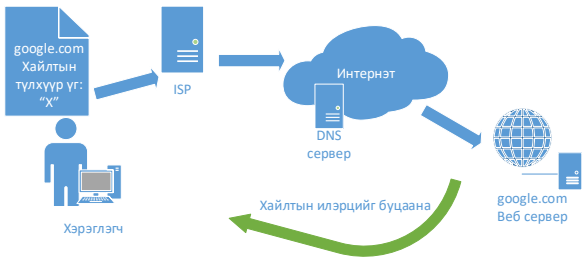
Б.Лхагва-Очир, Ж.Анхзаяа, Н.Угтахбаяр
* МУИС, ХШУИС, ЭХИТ
* blkhagvaa4810@gmail.com, ankhzaya@seas.num.edu.mn, ugtakhbayar@seas.num.edu.mn

Хураангуй— Tor сүлжээ нь интернэтэд байрлах компьютеруудаар, санамсаргүй дарааллаар, шифрлэгдсэн холболтоор дамжуулан хүссэн вэбийг хэрэглэгчид үзүүлдэг тул шууд холболтоос илүү ашуулгүйгээр холбогдох боломжтой юм. Ингэснээр Tor нь хэрэглэгчийн мэдээллийн нууцлалыг нэмэгдүүлж хамгаалах, мөн хаагдсан сайтууд руу орж үзэх боломж олгодог давуу талуудтай. Ер нь энгийн сүлжээний ашуулгүй байдлыг судалгаанд машин сургалтын аргыг хэрэглэх нь олон ч Tor сүлжээнд машин сургалтын аргыг хэрэглэсэн судалгаа тийм ч их биш байна. Тиймээс бид энэ удаагийн ажлаараа Tor сүлжээний нээлттэй өгөгдөл дээр машин сургалтын аргуудыг туршиж харьцуулан хамгийн сайн аргыг санал болгох зорилготой танилт хийж байна. Туршилтын үр дүнгээр CIC сангийн сүлжээний урсгалын хугацаанд суурилсан онцлогуудаар 22000 өгөгдөлд танилт хийж үзэхэд Бэйсийн төрлийн аргуудад 50-70% орчим зөв танилттай байхад нейрон сүлжээний төрлийн болон ойрын хөршийн аргуудад 85-99% танилттай байгаа нь өгөгдлийн түгэлт нормал биш байгаа нь харагдаж байгаа ба энэ төрлийн аргуудыг хэрэглэх санал болгож байна.

Түлхүүр үг— Сүлжээний ашуулгүй байдал, машин сургалт, Тор сүлжээ.

I. ОРШИЛ

Интернэт хэрэглэгчдийн хувийн мэдээллийн нууцлал, хамгаалалтыг хангах нь нэн тэргүүний, чухал асуудлуудын нэг болоод байгаа билээ. Засгийн газар, сэтгүүлчид, хакерууд зэрэг хөндлөнгийн этгээдүүд сүлжээний хэрэглэгчдийн хувийн харилцааг санаатай болон санамсаргүйгээр тагнах, чагнах үйл ажиллагаа их явуулдаг болсон. Тухайлбал сүлжээн дэх хэрэглэгчийн хандаж буй веб хуудас, хайлтын сайт дээр бичсэн түлхүүр үгс, хуваалцаж буй зураг, файл зэргийг хянаж, тагнах боломжтой байдаг. Ингэснээр тухайн хэрэглэгчийн хувьд мэдээллээ алдах эрсдэлтэй болж байгаа юм.

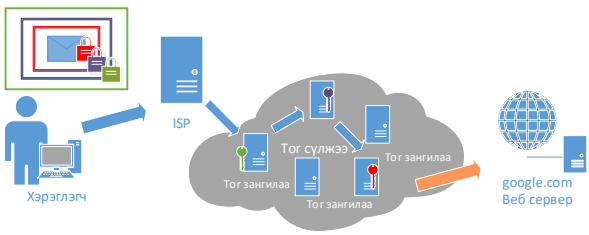


Зураг 1. Интернэтийн энгийн сүлжээ

Сүлжээний жирийн хандалтыг бид Зураг 1-д хэрэглэгч www.google.com сайтаас хайлт хийх, үр дүнг буцаах үйл явцыг харууллаа.

Хэрэглэгч www.google.com веб серверт холбогдох хүсэлтээ явуулахад DNS серверийн тусламжтайгаар тухайн веб серверийн IP хаягийг олж, сервер рүү пакетыг чиглүүлнэ. Сервер хүсэлтийг боловсруулж хэрэглэгчийн түлхүүр үгээр хайлтын үр дүнг хэрэглэгч рүү буцаадаг. Энэ үйл явцын турш хэрэглэгчийн IP хаягаар тухайн хэрэглэгчийн үйл ажиллагааг хянах боломжтой юм.

Харин Tor сүлжээн дэх Tor хөтөч буюу The Onion Router нь интернэт дэх хэрэглэгчдийн нууцлалтыг хангах зорилготой програм хангамж юм. Энэ нь хэрэглэгчийн нууцлалыг хангахдаа олон түвшинт кодлолыг ашигладаг. Tor програм хангамжийг компьютер дээрээ суулгасан хэрэглэгч веб сайт үзэх хүсэлт явуулахад тухайн пакет нь Tor зангилаануудтай холбогдох ба холболт бүрт өөр өөр кодлол ашиглаж хэрэглэгчийн нууцлал, хамгаалалтыг хэрэгжүүлдэг. Зураг 2-т Tor сүлжээнд хэрэглэгчийн нууцлалыг хангах процессыг үзүүлсэн байна.



Зураг 2. Tor сүлжээний процесс

Tor сүлжээнд дамжуулал хийж байгаа хэрэглэгчийн IP хаягийг нуух замаар хэрэглэгчийн хувийн мэдээллийг хамгаалдаг.

Tor сүлжээ нь хэрэглэгчийн хувийн мэдээллийн нууцлалыг сайжруулдаг давуу талтай ч хэд хэдэн дутагдалтай тал бий. Тухайлбал, хэрэглэгчийн мэдээлэл хамгийн багадаа гурван Tor зангилаагаар дамждаг учраас удаан, мөн Tor сүлжээний хэрэглэгчид хакерууд байх магадлал өндөр. Бид энэ ажлаараа хүлээж авах дамжуулах хугацааны утгуудыг гол онцлог болгон авч Tor төрөл мөн эсэхийг ангилах машин сургалтын аргуудыг туршин хамгийн сайн аргыг санал болгохыг зорилоо.

II. ХОЛБООТОЙ АЖЛУУД

Бидний ажилтай төстэй, хамааралтай ажлуудыг авч үзвэл дараах байдалтай байна. Бодит сүлжээний урсгалаас өгөгдлөө цуглуулан Mashael AlSabah нар сүлжээний траффикыг ангилан Тор сүлжээний гүйцэтгэлийг сайжруулах ажлыг танилцуулсан ба тэдний үр дүнгээр 200 орчим өгөгдөлд ангилалт хийхэд НейвБэйсийн арга танилт багатай ч Бэйсийн сүлжээ нь мөн л өндөр танилттай үр дүн үзүүлсэн байна.[1]

Мөн танилт хийгээгүй ч Тор сүлжээний хэрэглэгчийн нууцлал аюулгүй байдлыг сайжруулах шинэ боломж буюу олон Тор сүлжээний холимог хэлбэрийг санал болгосон ажлыг Amadou Moctar Kane нар хийжээ. [2]

Тор сүлжээн дэх вэб сайтын fingerprint-ийг сайжруулахын тулд Tao Wang нар 2 траффикийн ижил төстэй байдлыг хэмжих санаа арга санал болгосон ба тэдний санал болгож буй зайн хэмжигдэхүүний арга нь түгээмэл хэрэглэгддэг сангууд дээр тулгуур вектор машин, тэмдэгтийн цуваа харьцуулах зэргээс илүү үр дүн үзүүлсэн байна.[3]

Лондонгийн коллеж их сургуульд Мэдээллийн аюулгүй байдлын магистр хамгаалсан Otto Huhta судалгааны ажлаараа Тор сүлжээний олон шинжилгээ шийдвэрийн мод, санамсаргүй ой зэрэг аргуудыг ашиглах хийж олон сонирхолтой дүгнэлтэнд хүрсэн байна.[4]

Saman Feghhi нар судалгааны ажлаараа Тор сүлжээний Tor gateway клиент хоёрын дунд аттакыг илрүүлэх санааг дэвшүүлсэн ба тэдний ажлын үр дүнд к ойрын хөрш аргаар хугацааны мэдээлэл ашиглан этэрнэт сүлжээнд 90 хувь гаран, тор сүлжээнд 68 орчим хувьтай илрүүлсэн байна. [5]

Khalid Shahbar нар сүлжээний урсгалын анализыг мөн Тор сүлжээн дээр хийсэн бол Канадын их сургуулийн Arash Habibi Lashkari нар хугацааны мэдээллээр тор сүлжээний траффикыг ангилсан байна. [6,7]

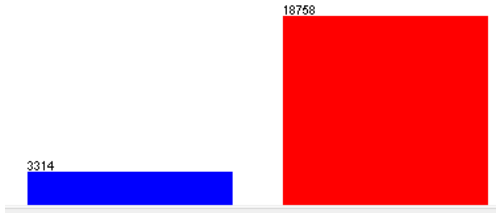
Олон улсад дээрх байдалтай судалгааны ажлууд олноо байгаа ч манай орны хувьд бидний судалгаагаар ховор байна. Тиймээс бид энэ удаагийн ажлаараа Тор сүлжээний өгөгдөл дээр машин сургалтын аргуудыг ашиглан танилт хийж, сайн аргыг санал болгохоор зорилоо.

III. МАШИН СУРГАЛТЫН АРГААР ТОР СҮЛЖЭЭГ ИЛРҮҮЛЭХ

Бид энэ удаагийн ажлаараа Канадын Нью Брунсвикийн их сургуулийн нээлттэй сангаас Тор сүлжээний шинжилгээний санг ашиглан машин сургалтын аргаар Тор мөн бишийг хугацааны утгуудаас хамааруулан ялгах ажлыг хийлээ.

Уг сан нь Тор төрлийн 3314 түүвэр, Тор биш төрлийн 18758 түүврийг агуулсан ба хугацааны 4 утгаар онцлог шинж болгон авсан. Сангийн мэдээллийг доорх зургаар харуулав.[7]

No.	Label	Count
1	TOR	3314
2	NONTOR	18758



Зураг 3. Өгөгдлийн сангийн мэдээлэл

Дээрх санг ашиглан машин сургалтын суурь аргууд дээр танилт хийж үзлээ. Туршилтын үр дүн хэсэгт хэрхэн ангилсан болон алдааны талаар тайлбарлана.

Машин сургалтын аргууд нь өөр өөрийн онцлог шинжүүдтэй тооцооллын хувьд ялгаатай ба үр дүн хэр сайн байх нь сургалтын өгөгдлийн түгэлтээс маш их хамаардаг. Тухайлбал статистик төрлийн аргууд нь өгөгдлийн Гауссын түгэлттэй гэж таамаглан дундаж, хазайлт зэрэг параметр тооцож, түүний утгаар шинэ өгөгдлийг ангилдаг. Хэрэв өгөгдлийн түгэлтэлт таамаглалын дагуу байвал үр дүн зөв сайн танина. Хэрэв статистик аргуудад үр дүн сайн биш байвал параметргүй аргуудыг хэрэглэх нь илүү байдаг.

IV. ТУРШИЛТЫН ҮР ДҮН

Бид машин сургалтын аргуудыг шалгахын тулд WEKA түүлийг ашиглан нийтлэг аргуудаас ангилалт хийж үзлээ. Нийт өгөгдлийн 66 хувийг сурган 33 хувийг таниулахад хамгийн сайн танилтын үр дүнг Нейрон сүлжээний арга үзүүлж байлаа.

Үр дүнгийн жишээ

Бэйсийн нетворк

Correctly Classified Instances	7440	99.1471 %
Incorrectly Classified Instances	64	0.8529 %

Олон давхаргат пресептрон

orrectly Classified Instances	6739	89.8054 %
Incorrectly Classified Instances	765	10.1946 %

RBF нетворк

Correctly Classified Instances	7271	96.895 %
Incorrectly Classified Instances	233	3.105 %

Энгийн логистик

Correctly Classified Instances	7448	99.2537 %
Incorrectly Classified Instances	56	0.7463 %

Хүснэгт 8 . Танилтын хувь

№	Машин сургалтын арга	Танилтын хувь
	Бэйсийн сүлжээ	99.14
	Олон давхаргат сүлжээ	89.80
	RBF сүлжээ	96.89
	Энгийн логистик сүлжээ	99.25

ДҮГНЭЛТ

Бид энэ удаагийн ажлаараа Тор сүлжээний шинжилгээнд машин сургалтын аргуудыг туршиж үзсэн ба хамгийн сайн танилтын хувийг Нейрон сүлжээний төрлийн аргууд нэлээд өндөр хувьтай зөв таньсан юм. Энэ нь бидний хэрэглэж буй өгөгдлийн түгэлт нормал түгэлттэй биш байгаа нь харагдаж байна. Энэ жишгээр бид өөрсдийн өгөгдлийн түгэлтийг амжилттай байгаа аргын үр дүнгээр тодорхойлох нь нэг чухал судалгаа юм. Өнөөдөр дэлхий нийтэд тор сүлжээний анализын олон ажлууд хийгдэж байгаа ба бид машин сургалтын аргуудыг ашиглан цөөхөн утгаар өндөр танилтанд хийж, олон төрлийн дүгнэлтийг хийж болох юм. Цаашид бид өөр олон онцлог шинжүүдийн хувьд олон төрлийн шинжилгээнүүдийг хийж боломжтой юм.

НОМЗҮЙ

[1] Mashael Alsabah, Kevin Bauer, and Ian Goldberg, Enhancing Tor’s Performance Using Real-Time Traffic Classification, In Proceedings of the 19th ACM Conference on Computer and Communications Security (CCS ’12). NY, 73–84. 2012.

[2] Kane A. M. Another Tor is possible Cryptology ePrint Archive 787. 2014.

[3] Khalid Shahbar A. Nur Zincir-Heywood, Traffic Flow Analysis of Tor Pluggable Transports, 11th Int. Conf. on Network and Service Management (CNSM), 178 – 181, 2015.

[4] Tor Pluggable Transports. [Online]. Available: https://www.torproject.org/docs/pluggable-transports.html.en

[5] Loesing, K. Safely collecting data to estimate the number of Tor users. https://lists.torproject.org/pipermail/tor-dev/2010-August/000467.html, 2012.

[6] Saman Feghhi, Douglas J. Leith, A First-Hop Traffic Analysis Attack Against Tor, In Communications (ICC), IEEE International Conference on, 1–6, 2015.

[7] Tao Wang and Ian Goldberg. Improved website fingerprinting on tor. In Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society, pages 201–212. ACM, 2013

[8] Habibi Lashkari A., Draper Gil G., Mamun M. and Ghorbani A. Characterization of Tor Traffic using Time based Features. In Proceedings of the 3rd International Conference on Information Systems Security and Privacy, Vol.1: ICISSP, pages 253-262, (2017). http://www.unb.ca/cic/research/datasets/tor.html