Нигеријска превара

Семинарски рад у оквиру курса
Рачунарство и друштво
Математички факултет

Павле Савић
pavlesavic1389@gmail.com

10. август 2021.

Сажетак

Превара 419 (такође позната и као нигеријска превара) је популаран облик преваре у којем преварант превари жртву да плати одређену суму новца под обећањем будуће, веће исплате. У овом раду обрадићемо начине организације ове преваре и како су они еволуирали кроз време, ослањајући се на рад француских истраживача са *Eurocom*-а [1]. Посебно говоримо о улози телефонских бројева као важних идентификатора за груписање порука и приказујемо начин на који преваранти воде своје кампање. Заправо, пошто жртва мора ступити у контакт са скамером, адреса е-поште и телефонски бројеви морају бити аутентични, ретко се мењају и користе се током дужег временског периода.

Садржај

1 Увод

Нигеријска превара [2] (превара 419) односи се на одељак 419 у нигеријском кривичном законику. Овај проблем познат је већ неколико деценија. Назив обухвата многе варијације ове врсте превара, попут превара са авансним провизијама, лажне лутрије, преваре са црним новцем итд. Првобитно је феномен преваре 419 започео путем поште, а затим се развио у посао који се прво водио путем факса, а касније путем е-поште. Кривично гоњење за такве криминалне активности је компликовано [3] и нападачи га често могу избећи. Због тога се извештаји о таквим илегалним радњама и даље појављују на друштевним мрежама и онлајн заједницама, нпр. сајт *419scam.org* [4] постоји како би се умањио ризик и помогло корисницима да идентификују преваре.

Данас, овај тип преваре често се сматра типом нежељене поште (енг. *spam*). Међутим, док се највећи део нежељене поште у новије време масовно шаље аутоматски коришћењем ботнета, нигеријска превара се и даље у великој мери обавља ручно. *Спамери* своје жртве нападају инжењерским напорима, док се *скамери* ослањају на људске факторе: сажаљење, похлепу и технике друштвеног инжењеринга користећи примитивније алате. Посебна карактеристика превара путем е-поште је комуникациони канал постављен да допре до жртве: *скамери* чешће користе е-пошту и/или телефонске бројеве [5] док је за друге облике *спама* вероватније да прослеђују своје жртве на контаминиране УРЛ-адресе. На пример, студија о *спам* кампањама из 2009. године [6] показује да 59% нежељене поште садржи УРЛ (*скам* је посматран као подскуп *спама*). Међутим, и поред огромне количине аутоматски генерисане нежељене поште, 419 *скам* поруке и даље представљају сталан проблем који узрокује значајне личне финансијске губитке за бројне жртве широм света.

Традиционални *спам* и *скам* сценарији (не 419) темељно су проучени (нпр. [6, 7]), већина техника ове природе ослања се на огромне количине сличних порука. Насупрот томе, 419 поруке вероватније ће бити послате у мањем броју копија и са налога веб поште. На тај начин, нападачи имају за циљ да остану неопажени од стране традиционалних филтера за нежељену пошту и избегну скретање пажње на злоупотребљене налоге веб поште. Прецизне методе дистрибуције 419 *скам* порука нису проучаване тако детаљно као, на пример, дистрибуција *спам* поште путем *ботнета*. На основу *Microsoft Security Intelligence Reports* [8], 419 поруке чиниле су у просеку 8% нежељене е-поште у периоду 2009-2014.

Студија из 2013. године [5] описује употребу телефонских бројева у бројним злонамерним активностима. Аутори су показали да су телефонски бројеви које користе *скамери* често активни током дужег временског периода и да се често користе изнова и изнова у различитим мејловима, што их чини потенцијално атрактивном метом за повезивање *скам* порука и идентификацију преваре. Ову хипотезу подробније ћемо тестирати користећи телефонске бројеве и друге функције е-поште за аутоматско откривање и проучавање *скам* кампањи коришћењем јавног скупа података.

Наш циљ је да проучимо како *скамери* организују своје преваре, гледајући међусобне везе између налога е-поште, телефонских бројева и тема е-поште које користе. У ту сврху користи се алгоритам одлучивања са више критеријума за ефикасно *кластеровање* (груписање) *скам* порука које деле минималан број заједничких особина,

чак и у присуству променљивих (енг. *volatile*) особина. Због овог скупа заједничких карактеристика, *скам* мејлови који потичу од истих нападача вероватно ће бити груписани заједно, што нам омогућава да стекнемо увид у *скам* кампању. Поред тога, оцењујемо квалитет и доследност наших резултата груписања. Ради тога вршимо анализу прага осетљивости, као и процену хомогености кластера користећи компактност графикона и енг. *Adjusted Rand Index* као метрике.

У анализи приказаној у обрађеном раду [1] идентификовано је преко 1000 различитих превара, и за већину њих, телефонски бројеви представљају камен темељац који је омогућио повезивање различитих делова. На овај начин откривене су и неке кампање већег обима (енг. *macro-cluster*), сачињене од међусобно слабо повезаних *скам кластера* који одражавају различите операције истих превараната. То се може приписати различитим *скамовима* које организују исте криминалне групе, јер проналазимо исте телефонске бројеве или налоге е-поште који се поновно користе у различитим подкампањама.

Експериментално је утврђено да се ове методе могу искористити за проактивно идентификовање нових превара (или варијанти претходних) брзим повезивањем нове преваре са текућим кампањама, што би могло олакшати рад правним институцијама приликом кривичног гоњења превараната. Овај приступ се такође може искористити и за побољшање истрага других шема сајбера криминала евидентирањем и истраживањем различитих група сајбер криминалаца на основу њихових активности на мрежи. Употребљивост ове методологије је већ показана у контексту других безбедносних истрага, као што су анализе *спам* поште дистрибуиране путем *ботнета* [9], циљаних напада [10] и лажних антивируса [11].

2 Скуп података

У овом одељку описујемо скуп података коришћен за анализу 419 превара и пружамо неке статистичке податке о тим кампањама. Постоје различити извори превара које корисници често пријављују које потом сакупљају наменске заједнице, форуми и друге онлајн групе. Подаци одабрани за анализу су прикупљени са *419scam.org - a 419 scam aggregator* јер овај сајт пружа обиман скуп унапред претпроцесираних података: заглавља, тела е-поште и неке већ издвојене атрибуте е-поште, попут категорије *скама* и телефонских бројева. Важно је нагласити да подаци о IP-адресама недостају. Анализирана је пошта за период од јануара 2009. до августа 2012. године.

У овој студији такође је искоришћена чињеница да телефонски бројеви могу означавати географску локацију, обично земљу у којој је телефон регистрован. Иако то не доказује са сигурношћу порекло поруке или *скамера*, ипак реферише на земљу где се превара спроводи, а такође и повећава поверење жртве у примљену поруку. На пример, примање партнерске понуде из Уједињеног Краљевства могло би изгледати сумњиво ако телефонски контакт има нигеријски префикс или обавештење о лутрији са контакт подацима из афричке земље, док је жртва из Европе. Штавише, као што је показано у студији [5] бројеви мобилних телефона коришћених за нигеријску превару прецизно указују на државу пребивалишта власника телефона (нападача) јер је пронађено неколико случајева роминга. Због тога је атрибут телефонског броја довољно прецизан да означи географско

порекло напада.

Добијени скуп података састоји се од 36.761 порука са 11.768 јединствених телефонских бројева. Опште статистике података приказане су у табели 1. Прво што треба приметити је да је број адреса е-поште три пута већи од броја телефонских бројева, што указује на могућност набавке поштанских сандучића у злонамерне сврхе. Међутим, однос је и даље прилично низак, што указује на прилично јефтин и лак приступ телефонским бројевима. Још једна специфичност овог скупа података је да свака порука може садржати неколико адреса е-поште и телефонских бројева, при чему број различитих адреса е-поште може бити чак 5 по поруци: *from* адреса,*reply* адреса и друге адресе наведене у тексту поруке. Дакле, иако је прикупљено 36.761 порука, из њих је извучено 112.961 адреса е-поште.

Табела 1: Опште статистике

Опис	**Број**
Скам поруке	36.761
Јединствене поруке	26.250
Укупно мејл адреса	112.961
Јединствене мејл адресе	34.723
Укупно телефонских бројева	41.320
Јединствени телефонски бројеви	11.768
Број држава	12

У обрађеном скупу података нису примећени значајнији налети *скам* порука током трогодишњег периода (на месечном нивоу), што указује на то да су се поруке е-поште константно дистрибуирале током времена. Такође је важно напоменути да је скуп података углавном ограничен на европске и афричке регионе (са само неколико азијских узорака), што је последица начина на који власници вебсајтова прикупљају и класификују податке. Ипак, географска распрострањеност поменутих континената одражава се и на посматрани скуп података, искључујући малобројне изузетке.

Ради бољег разумевања скупа података, треба погледати време током којег су *скамери* у *скам* порукама користили мејл адресе и телефоне. 71% адреса е-поште у овом скупу података коришћено је само током једног дана. Преостале су коришћене у просеку по 79 дана. Телефонски бројеви имају дужи век трајања: 51% коришћен је само 1 дан; остатак је коришћен у просеку 174 дана (око 6 месеци), што га чини важном особином приликом анализа засноване на *кластеровању*.

Табела 2 резимира географску дистрибуцију телефонских бројева. Бројеви из Уједињеног Краљевства двоструко су чешћи од нигеријских, и три пута чешћи од оних из Бенина, треће највеће групе. Холандија и Шпанија водеће су земље у Европи. УК треба сматрати посебним случајем. По извештајима *419scam.org* и [5], сви бројеви телефона из УК у овом скупу података припадају енг. *personal numbering services* - сервисима који се користе за прослеђивање телефонских позива на друге телефонске бројеве и служе за маскирање стварне дестинације позиваоца. У посматраном скупу података постоји 44% таквих телефонских бројева (сви са префиксом УК), још 44% су бројеви мобилних телефона, 12% су фиксне линије [5], мање од 1% телефона су непостојећи.

Табела 2: Телефони по државама

Држава	**Укупно телефона**	**Укупно (%)**
Уједињено Краљевство	4.499	43
Нигерија	3.121	30
Бенин	1.448	14
Јужна Африка	562	5
Шпанија	372	4
Холандија	263	3
Обала Слоноваче	89	1
Кина	68	1
Сенегал	47	0.5
Того	11	0.1
Индонезија	1	0.01

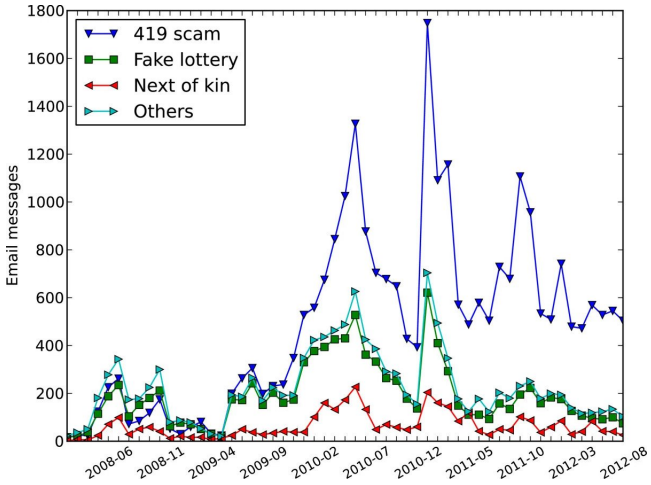
Порукама које су иницијално класификоване као превара 419 *419scam.org* [4] додељује и категорију. Око 64% тих порука сврстано је у категорију *'419 scam'* која је поткатегорија преваре 419 и односи се посебно на врсте финансијских превара, нпр. преваре са трансакцијама, изгубљена средства итд. Према извештају [4] већина преосталих мејлова (24%) припада поткатегорији *'Fake lottery'*. Међутим, ова дистрибуција се временом мењала као што је приказано на слици 1. Посебно, велика разлика може се приметити између 2009. и 2011. када је *'419 scam'* категорија постала доминантна. У августу 2012. било је 5 пута више *'419 scam'* него *'Fake lottery'* порука. То може бити последица застарелог процеса категоризације, јер се *скам* може мењати и еволуирати временом. Из тог разлога, користи се процес аутоматског идентификовања врсте преваре који се заснива на фреквенцији речи у порукама. Такође примећујемо да је већина *'Fake lottery'* превара повезана са европским телефонским бројевима што сугерише да се ова категорија шаље циљаној публици. Код *'419 scam'* порука, нападачи користе готово подједнако често нигеријске и британске бројеве (слика 2). Од 2011. значајнију улогу почиње да игра и Бенин.

3 Методе анализе података

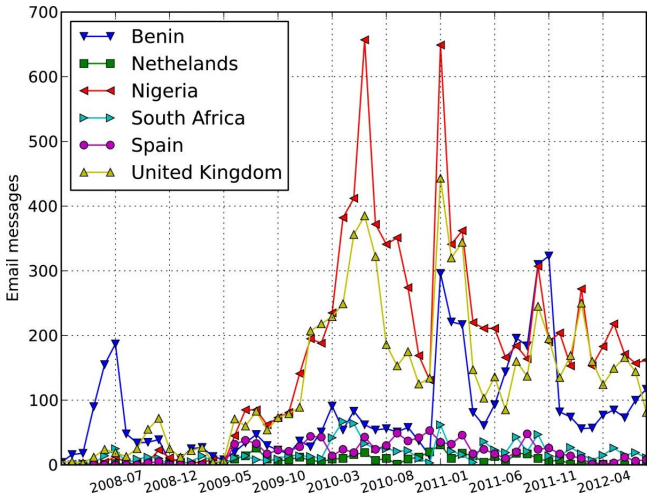
У овом одељку описујемо методе које се за користе за идентификовање група сличних 419 *скам* порука за које се верује да су део истих кампања, а потом представљамо резултате. Користимо две различите метрике за процену квалитета и доследности креираних кластера (кампања). Коначно, из тела ових порука издвајамо кључне речи које се највише понављају како бисмо побољшали категоризацију.

3.1 Кластеровање *скам* порука : *'TRIAGE'* приступ

Да бисмо идентификовали групе *скам* порука које имају изгледа да буду део кампање коју је организовала иста група људи, кластеровали смо све *скам* поруке користећи *'TRIAGE' - framework* за безбедносно истраживање података који користи предности анализе више



Слика 1: Scam email categories over time

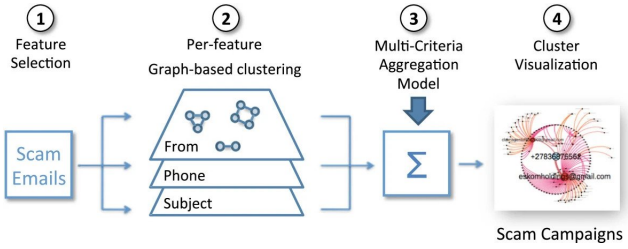


Слика 2: '419 scam' category phone numbers over time by countries

критеријума за груписање догађаја на основу подскупова заједничких елемената (касније званих особинама). Захваљујући овом приступу кластеровања по више критеријума, *'TRIAGE'* идентификује сложене обрасце у подацима, откривајући понекад варирајуће односе између низа повезаних или различитих догађаја. *'TRIAGE'* се најбоље може описати као сигурносни алат дизајниран за извлачење обавештајних података који помаже у одређивању образаца понашања нападача (тј. тактика, техника и процедура или ТТП-ова), истичући *како* делују, а не *шта* раде. Овај *framework* [12] је већ показао своју употребљивост у контексту других безбедносних истрага, на пример, лажних АВ кампања [11], *спам* ботнетова [9] и циљаних напада [10].

Слика 3 приказује ток рада *'TRIAGE'-а* примењеног на посматрани скуп података. У првом кораку се одабирају карактеристике

е-поште које се дефинишу као критеријуми одлучивања за повезивање мејлова попут адресе е-поште пошиљаоца (од), наслова поруке, датума слања, адресе одговора (у заглављу мејла), броја телефона и било које друге адресе е-поште која се налази у самој поруци (телу е-поште). У другом кораку, *'TRIAGE'-a* граде се односи међу свим узорцима у односу на изабране особине користећи одговарајуће метрике сличности. Прецизније, коришћене су различите мере сличности оријентисане на стрингове, као што су Левенштајнова сличност (за *subject*) и сличност H-грама (за *from*, *reply*, *email body*) [13]. За карактеристике као што су *phone* и *date* коришћена је једноставна метода поређења једнакости - најприкладније одговара семантици ових особина у конкретном случају.



Слика 3: 'TRIAGE' workflow on scam dataset

У трећем кораку, појединачне сличности карактеристика се агрегирају помоћу агрегатног модела који одражава понашање високог нивоа које је дефинисао аналитичар, који може одредити, на пример да је потребно најмање k веома сличних особина (од n) да бисмо узорке могли приписати истој кампањи. Омогућено је додељивање тежина тако да одређеним карактеристикама можемо бити дат већи или мањи значај. У табели 3 приказан је скуп тежина коришћен на нашем скупу података, у којем је наглашен значај телефонских бројева и тема е-поште. Особинама везаним за е-адресе пошиљаоца дат је средњи значај, док је датуму слања дат мали значај. Укупна добијена вредност (агрегирана вредност) се потом користи као улаз за класичне *graph clustering* алгоритме, попут *minimal cut* и *connected components* који нам могу дати кластере произвољне величине и облика.

Табела 3: Weights of individual features (total = 1)

Feature	**Importance**
Phone	0.30
From	0.12
Reply	0.18
Subject	0.25
Email body	0.1
Date	0.05

У овој фази, важно је нагласити да се, осим можда телефонског броја који користи нападач (осим ако је овај број лажан), ниједна друга особина укључена у анализу не може сматрати сама по себи деовољном за приписивање *скам* порука истом нападачу. На пример,

чињеница да је иста (или слична) адреса е-поште пошиљаоца коришћена у две *скам* поруке, чак и послате истог датума, не мора нужно значити да те две поруке потичу од истог појединца. Само једна или две заједничке карактеристике могу се појавити као последица случајности, или можда неке уобичајене технике коју нападачи деле. Ово мотивише избор фузионог модела са више критеријума који укључује различите тежине дефинисане према *Regular Increasing Monotone - RIM* квантификатору [14], што нам омогућава да моделирамо стратегије као што су - најмање k јаких корелација је потребно за повезивање два мејла.

3.2 Резултати кластеровања и експериментална валидација

Алат за кластеровање *'TRIAGE'* идентификовао је 1.040 кластера који се састоје од најмање 5 *скам* порука повезаних различитим комбинацијама особина. Због начина генерисања ових кластера (тј. агрегације по више критеријума), очекујемо да ти кластери мејлова представљају различите кампање, које потенцијално могу организовати исти појединци - пошто мејлови у истом кластеру деле неколико заједничких особина.

Табела 4 пружа неке глобалне статистике израчунате у 250 највећих *скам* кампањи. У више од половине ових кампања преваранти користе само два различита телефонска броја, али и даље користе више од пет различитих поштанских сандучића како би добили одговоре од својих жртава. Већина кампања је прилично дуготрајна (у просеку траје око годину дана). Напомињемо да су величине кластера у просеку мале, што указује на то да постоји много малих, изолованих кампањи и да само неколико десетина порука припада истој кампањи. Ово би такође могла бити последица процеса прикупљања података; ипак, предвиђамо да би ово такође могло одражавати понашање *скамера* који би можда желели да остану *испод радара*. Заиста, велике количине истих порука е-поште имале би већи потенцијал да угрозе њихове *скам* операције, јер би то постало видљиво за филтере нежељене поште засноване на садржају (енг. *content-based filters*), па би бивали блокирани у ранијим фазама филтрирања е-поште.

Табела 4: Global statistics of the top 250 clusters

Statistic	**Average**	**Median**	**Maximum**
Number of emails	38	28	376
Number of from	13.9	9	181
Number of replies	6.2	5	56
Number of subjects	9.9	7	114
Number of phones	2.5	2	34
Duration (in days)	396	340	1.454
Number od dates(distinct)	27.9	22	259
Compactness	2.5	2.4	5.0

3.2.1 Процена резултата кластеровања

Како је кластеровање података у основи приступ класификације без надзора, важно је проценити резултате груписања помоћу објективних критеријума, како би се потврдила ваљаност резултата. Постоје два приступа за вршење ове валидације: **екстерна** и **интерна** евалуација.

Да бисмо проценили квалитет наших резултата груписања, испитана је њихова укупна компактност, разврстана по појединачним карактеристикама. Компактност графикона је индекс ваљаности кластера који показује колико су кластери *компактни* (или хомогени), на основу њихових карактеристика унутар повезивања. Компактност одражава просечну сличност ивица између два објекта кластера.

Пошто алат *'TRIAGE'* прати све појединачне везе у графиконима сличности, такође је могуће израчунати удео порука е-поште које су повезане посебним комбинацијама особина унутар кластера. Ово може бити веома корисно за разумевање разлога за формирање кластера и на тај начин пружити увид у *стабилне* (мање променљиве) карактеристике које користе преваранти приликом извођења нових кампања.

У табели 5 можемо приметити да најчешћа комбинација која повезује *скам* поруке (у 13% случајева) укључује телефонски број, тему, као и све три адресе е-поште (пошиљалац, одговор, тело мејла). Да бисмо потврдили интуицију о важности одређених карактеристика (телефонских бројева, и у мањој мери, адреса е-поште) и њиховој ефективној улози у идентификовању кампањи, разматрамо све везе засноване на сличности унутар кластера. Уочавамо да особине које су углавном одговорне за повезивање *скам* порука у кластерима укључују телефонске бројеве (у 88% случајева), затим адресу е-поште одговора (за 66% веза). Није изненађујуће што се адреса пошиљаоца (која се лако може лажирати) мења много чешће и користи се као повезујућа карактеристика у само 46% случајева.

Табела 5: Top coalitions of features across all clusters

Coalition	**Percentage (%)**
(phone, subject, from, reply, email body)	13
(phone, reply, email body)	12
(phone, subject, reply, email body)	11
(phone, from, reply, email body)	7
(phone, subject)	6
(phone, from)	5
(phone, reply)	4
(phone, reply, subject)	4
(phone, reply, subject, from)	4
others	33

4 *Макро-кластери*: повезивање поткампањи

У следећем кораку смо *скам* кампање посматрали из шире перспективе: тражењем слабо повезаних кластера. Циљ је био да се укаже на могуће кампање већег обима, које су сачињене од међусобно слабо повезаних *скам* операција (тј. различитих превара). У ту сврху користили смо само адресе е-поште и телефонске бројеве, будући да се други атрибути не сматрају подацима за личну идентификацију. У ствари, тражили смо кластере који деле најмање једну адресу е-поште и/или телефонски број и користили те информације за изградњу такозваних *макро-кластера*.

Као резултат тога, идентификовали смо скуп од 845 изолованих кластера и други скуп од 195 повезаних кластера, при чему се потоњи састоји од 62 *макро-кластера*. Карактеристике првих шест макро кампања приказане су у табели 6. Ови *макро-кластери* су посебно занимљиви јер се састоје од скупа *скам* кампањи које су слабо повезане и стога могу бити организоване од стране истих сајбер криминалаца. Заправо, везе између различитих кластера алгоритам за груписање је сматрао преслабим, због шеме одлучивања и прагова постављених за параметре, па су стога ти *скамови* на крају груписани у засебне кластере. Међутим, ове слабе везе се лако могу повратити, а на аналитичару је да истражи колико су те међусобне везе заиста значајне. Заиста, сајбер истражитељу много је лакше да почне од скупа заиста смислених кластера, а затим да постепено повећава прагове до тачке у којој може сам одлучити да престане са спајањем група података, јер више нема смисла приписивати различите кампање истој групи због недостатка доказа.

Табела 6: Macro-cluster, mean values of attributes

Number	Number of campaigns	Phones	Mailboxes	Subjects	Duration (years)	Countries	Topics
1	14	44	677	223	4	4	Lottery, lost funds, investments
2	43	163	1.127	463	4	7	Lottery, banks, diplomats, FBI
3	6	18	128	80	4	4	Lottery
4	5	8	111	51	3.5	2	Packaging, lottery, loans
5	6	7	201	96	1	1	Lottery, UPS & WU delivery
6	4	7	82	33	2	1	Diplomats, monetary and payments scam

5 Закључак

У овом раду анализирана је студија француских истраживача у којој је емпиријски показано постојање 419 преваре и кључна улога коју телефонски бројеви и адресе е-поште имају у овим преварама, за разлику од других шема сајбер криминала у којима се адресе е-поште често могу лажирати, а телефонски бројеви ретко користити. Помоћу технике вишедимензионалног кластеровања идентификовано је на посматраном скупу података око 1.000 *скам* кампањи.

Приказани су начини на које се кластери издвајају и њихове опште статистике. На крају, указано је на методе које нам омогућују откривање постојање слабих веза између наизглед неповезаних сајбер превара које чине макрокампање за које са великом вероватноћом можемо тврдити да су вођене од стране истих сајбер-криминалаца. Утврђено је да су неке од ових макро-кампањи географски распоређене у

неколико земаља, афричких и европских. Анализа је открила велику разноликост у начину увлачења жртве у превару, као и у вођењу кампање. Међу *скамерима* постоји велика конкуренција у актуелности тема преваре.

Литература

[1] Isacenkova, Jelena, et al. "Inside the scam jungle: A closer look at 419 scam email operations."EURASIP Journal on Information Security 2014.1 (2014): 1-18.

[2] Definition of Nigerian Scam, http://www.419scam.org/419scam.htm

[3] Buchanan J, Grant AJ: Investigating and prosecuting Nigerian Fraud. United States Attorneys' Bulletin 2001., 49(6):

[4] 419 Scam - fake lottery Fraud phone directory, http://www.419scam.org/419-by-phone.htm

[5] Costin A, Isacenkova J, Balduzzi M, Francillon A, Balzarotti D: The role of phone numbers in understanding cyber-crime. In 11th International Conference on Privacy, Security and Trust (PST 2013). Tarragona, Catalonia; 10-12 July, 2013.

[6] Pathak A, Qian F, Hu YC, Mao ZM, Ranjan S: Botnet spam campaigns can be long lasting: evidence, implications, and analysis. In SIGMETRICS '09 Proceedings of the Eleventh International Joint Conference on Measurement and Modeling of Computer Systems Seattle, WA, 15-19 June. New York: ACM; 2009:13-24.

[7] Anderson DS, Fleizach C, Savage S, Voelker GM: Spamscatter: Characterizing internet scam hosting infrastructure. PhD thesis, University of California, San Diego 2007

[8] Microsoft Security Intelligence Report (2008-2012), http://www.microsoft.com/security/sir/archive/default.aspx

[9] Thonnard O, Dacier M: A strategic analysis of spam botnets operations. In Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, CEAS '11. New York: ACM; 2011:162-171.

[10] Thonnard O, Bilge L, O'Gorman G, Kiernan S, Lee M: Industrial espionage and targeted attacks: understanding the characteristics of an escalating threat. In RAID, Amsterdam, The Netherlands, 12-14 September. New York: Springer; 2012:64-85.

[11] Cova M, Leita C, Thonnard O, Keromytis AD, Dacier M: An analysis of rogue AV campaigns. In Proceedings of the 13th International Conference on Recent Advances in Intrusion Detection. RAID'10, Berlin, Heidelberg: Springer-Verlag; 2010:442-463. [http://portal.acm.org/citation.cfm?id=1894166.1894196]

[12] Thonnard O: A multi-criteria clustering approach to support attack attribution in cyberspace. PhD thesis, École Doctorale d'Informatique, Télécommunications et Électronique de Paris, 2010

[13] Kondrak G: N-gram similarity and distance. In Proceedings of the 12th Conference on String Processing and Information Retrieval Buenos Aires,2–4November. Heidelberg: Springer-Verlag Berlin; 2005:115-126.

[14] Zadeh LA: A computational approach to fuzzy quantifiers in natural languages. Comput. & Math. Appl 1983, 9: 149-184.