

**FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO  
PORTO**

# **Towards Mitigating Unwanted Calls in Voice Over IP**

**Muhammad Ajmal Azad**



**FEUP** FACULDADE DE ENGENHARIA  
UNIVERSIDADE DO PORTO

Programa Doutoral em Engenharia Electrotécnica e de Computadores

Supervisor: Ricardo Santos Morla

June 2016

© Muhammad Ajmal Azad, 2016

# **Towards Mitigating Unwanted Calls in Voice Over IP**

**Muhammad Ajmal Azad**

Programa Doutoral em Engenharia Electrotécnica e de  
Computadores

*I Dedicate This Thesis To My Parents and Wife*

For their endless love, support and encouragement.

# Acknowledgments

First and foremost, I would like to express my special gratitude and thanks to my advisor, Professor Dr. Ricardo Santos Morla for his continuous support, supervision and time. His suggestions, advice and criticism on my work have helped me a lot from finding a problem, design a solution and analyzing the solution. I am forever grateful to Dr. Morla for mentoring and helping me throughout the course of my doctoral research..

I would like to thanks my friends Dr. Arif Ur Rahman and Dr. Farhan Riaz for helping in understanding various aspects of research at the start of my Ph.D, Asif Mohammad for helping me in coding with Java, and Bilal Hussain for constructive debate other than academic research and continuous encouragements in the last three years.

Of course acknowledgments are incomplete without thanking my parents, family members and loved ones. I am very thankful to my parents for spending on my education despite limited resources. They taught me about hard work, make me to study whenever I run away, encourage me to achieve the goals, self-respect and always encourage me for doing what i want. I am thankful to my sisters and brothers for continuous support, prayers and encouragements. I am proud of you all and love you all very much. Most importantly, my sincere thanks goes to my beloved wife who was always there whenever I need her. She single handedly cared our two daughters and allows me to concentrate on my research. I also thank little daughters Mehar Fatima and Haniya Hussain for coming into my life and thereby bringing a lot of happiness for me. I am also thankful to my sisters (Nayyab,Seemab and Eman) for their prayers and support, and my uncle Prof. Dr. Razzaque H. Bhatti and his family for invaluable support during my bachelor and master studies. Finally, i am very grateful to all my family members, friends, and brothers who

supported me directly and indirectly during this PhD work.

I am thankful to Renata Rodrigues at INESC TEC and the staff at Faculty of Engineering for making arrangement for my conferences. I am also thankful to FCT (Portuguese national funding agency for science, research and technology), FEUP (Faculty of Engineering University of Porto, Portugal) and NeTs project for providing me funding for my PhD.

Muhammad Ajmal Azad

# Resumo

A tecnologia VoIP (Voice over Internet Protocol) permite a realização de chamadas telefónicas baratas através da Internet. Uma vez que a tecnologia VoIP utiliza a mesma infraestrutura de Internet para o transporte da sinalização e da voz, está sujeita a todas as ameaças de segurança que afetam a Internet. Uma dessas ameaças é o spam de voz (em VoIP chamado SPIT), que é semelhante ao spam de e-mail mas que tem consequências mais severas do que o spam de e-mail porque uma chamada de voz requer uma resposta em tempo real do destinatário da chamada. De modo a aumentar a produtividade dos utilizadores desta tecnologia e prevenir perdas devido à fraude a ao spam, é extremamente importante identificar e bloquear o spam antes que afete e desagrade um número potencialmente elevado de utilizadores da tecnologia.

O desafio no desenho de sistemas autónomos de deteção SPIT é a utilização em simultâneo de características da chamada e da rede social que sejam difíceis de contornar pelos spammers. Para endereçar este desafio, esta tese apresenta um sistema de deteção autónomo de SPIT chamado Caller-REP que consiste em dois módulos. 1) Um módulo de reputação – que calcula a reputação do utilizador através da utilização simultânea da duração das chamadas do utilizador, do débito de chamadas do utilizador, e do número de destinatários únicos que o utilizador chama. O cálculo da reputação desta forma poderá atribuir valores de reputação pequenos aos spammers devido às suas redes de chamadas desequilibradas e valores de reputação grandes para os utilizadores legítimos. 2) Um módulo de deteção – que, utilizando os valores de reputação, calcula de forma automática um limiar abaixo do qual o utilizador é classificado como spammer.

Spammers e geradores de publicidade à distância têm como alvo um grande número de destinatários que estão geralmente distribuídos por vários fornecedores de serviço (SPs). Os sistemas de deteção autónomos consideram informação dos utilizadores registada localmente para distinguir spammers de utilizadores legítimos. Parece óbvio que a colaboração entre SPs poderá melhorar a exactidão e o tempo de deteção mas isto depende da quantidade de informação partilhada entre SPs. Os SPs partilham informação com outros SPs de forma relutante, tipicamente porque são concorrentes e porque estão preocupados com a privacidade dos seus clientes e da sua operação. Uma abordagem para convencer os SPs é a de partilha apenas de informação summarizada e com um sistema central confiável para proteger a sua privacidade. Para atingir o objetivo de colaboração atenta à privacidade entre SPs, esta tese prope COSDS (Collaborative SPIT detection System, sistema de deteção de SPIT colaborativo) que requer a colaboração entre SPs e um repositório centralizado através da partilha de valores de reputação. O repositório centralizado (CR) calcula a reputação global (GR) dos utilizadores através da agregação dos valores de reputação local e responde aos SPs com a decisão e com os valores de GR. Um adversário no CR estaria numa posição mais difícil para obter informação privada sobre os utilizadores e os fornecedores de serviço. Spammers e geradores de publicidade à distância poderão ter identidades de chamada múltiplas para contornar o sistema de deteção de SPIT. Estabelecer uma ligação entre identidades que pertencem a uma mesma pessoa física é importante para a identificação mais atempada de spammers e para a caracterização completa do comportamento de utilizadores legítimos com mais de uma identidade de chamada. O desafio no que diz respeito a isto tem duas partes: a ligação de identidades e o cálculo de valores de reputação para cada indivíduo através da combinação de informação das suas várias identidades. Para endereçar este desafio, esta tese apresenta um sistema chamado EIS (Early Identification of Spammer, identificação atempada de spammer) que utiliza características da rede social e das chamadas para interligar identidades semelhantes que pertençam ao mesmo indivíduo. A reputação é então calculada para o indivíduo em vez

da identidade de chamada e o indivíduo classificado como spammer se a sua reputação estiver abaixo do limiar calculado automaticamente. A ligação de identidades de chamada poderá não só ser útil na deteção de spammers que mudam frequentemente de identidades de chamada mas também poderá ajudar na deteção de redes criminais. As abordagens propostas nesta tese podem, juntas, ter um impacto significativo na identificação atempada de spammers numa rede VoIP sem serem intrusivas para o utilizador nem precisarem de mudanças na semântica e na arquitetura da rede VoIP.

# **Abstract**

Voice over Internet Protocol (VoIP) is the technology that allows people to make cheap telephone calls over the Internet. As VoIP uses the same Internet infrastructure for the transport of signaling and voice, it is subject to all security threats already effecting the Internet. One such threat is voice spam (termed as SPIT in VoIP), which is similar to e-mail spam but has more severe consequences than the email spam because voice call requires real-time response from the call recipient. In order to increase productivity of users of this technology and preventing losses due to fraud and spamming, it is extremely important to identify and block spam before it affects and displease a potentially large number of users of the technology.

The challenge in a design of standalone SPIT detection system is to simultaneously use call and social network features that are difficult to be circumvented by spammers. To address this challenge, this thesis presents the standalone SPIT detection system called Caller-REP that consists of two modules. 1) A reputation module – that computes reputation of the user by collectively using call duration of the user, call-rate of the user and total number of unique recipients the user called. The computation of reputation in this way would assign small reputation scores to spammers because of their unbalanced calling networks and high reputation scores to legitimate users. 2) A detection module – that computes automated threshold using reputation scores below which the user is classified as a spammer.

Spammers and telemarketers target a very large number of recipients usually dispersed across many Service Providers (SPs).The standalone detection systems consider locally

recorded information of users while differentiating spammers from the legitimate users, thus prolong detecting spammers. Obviously, collaboration among SPs would improve the detection accuracy and detection time, but this depends on the amount of information shared between SPs. SPs are reluctant in exchanging information to other SPs because they are business competitor and are worried about privacy of their customers and their operational data. SPs can be convinced with the exchange of summarized information to the centralized trusted system so to protect their privacy. To achieve the objective of privacy-aware collaboration among SPs, this thesis proposes COSDS (Collaborative SPIT detection System) that require collaboration between SP and the centralized repository with the exchange of reputation scores. The Centralized Repository (CR) computes global reputation (GR) of users by aggregating their local reputation scores and responds back SPs with decision and GR scores. The adversary on the CR would be in a more difficult position to obtain private information about the users and service providers.

Spammers and telemarketers would have multiple calling identities to circumvent the SPIT detection system. The linking of identities that belongs to one physical user is important for early identification of spammers and for characterizing the complete behavior of legitimate users having more than one calling identity. The challenge in this regard is twofold: first linking of identities and secondly computation of reputation scores for individual by combing information from all his identities. To address this challenge, thesis presents a system called EIS (Early identification of Spammer) that uses social network and call features for connecting similar identities that belong to one individual. The reputation is then computed for the individual rather than for the identity and individual is classified as spammer if his reputation is less than automated threshold. The identity linking would not only help in early detection of spammer frequently change identities but would also provide effectiveness in detecting criminal rings. All approaches proposed in this thesis can together have a significant impact on the early identification of spammer in a VoIP network without being intrusive to end-users and

without requiring any change in the VoIP network semantics and architecture.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Spam Over Internet Telephony . . . . .	2
1.1.1	Consequences of Spam over Internet Telephony . . . . .	3
1.1.2	Motivation of VoIP Spammer . . . . .	4
1.1.3	SPIT Differences from E-mail Spam . . . . .	5
1.1.4	Why is it Hard to Detect SPIT Caller ? . . . . .	7
1.2	Behavior-based Collaborative SPIT Detection System . . . . .	9
1.3	Thesis Contributions . . . . .	13
1.4	Publications . . . . .	16
1.5	Thesis Structure . . . . .	17
<b>2</b>	<b>Systems for Detecting Unwanted Communications in a VoIP Network</b>	<b>18</b>
2.1	Stand-alone SPIT detection . . . . .	19
2.1.1	Content-based Detection Systems . . . . .	19
2.1.2	Challenge/Response-based Detection Systems . . . . .	21
2.1.3	Access List-based Detection Systems . . . . .	23
2.1.4	Cost-based Detection Systems . . . . .	24
2.1.5	Policy-based Detection Systems . . . . .	24
2.1.6	Legislation-based Detection Systems . . . . .	25
2.1.7	Multi-Stage Detection Systems . . . . .	25
2.1.8	Call Statistics-based Detection Systems . . . . .	26
2.1.9	Device Fingerprinting-based Detection Systems . . . . .	27
2.1.10	Honeypot-based Detection Systems . . . . .	27
2.1.11	Reputation-based Detection Systems . . . . .	28
2.2	Collaborative Detection Systems . . . . .	30
2.2.1	Collaborative SPIT Detection . . . . .	31
2.2.2	Collaborative Spam Detection in Email Network . . . . .	32
2.3	Identity Linking . . . . .	33

2.4	Discussion . . . . .	36
<b>3</b>	<b>SPIT Detection System Based on Social Reputation</b>	<b>39</b>
3.1	VoIP (Voice over IP) . . . . .	39
3.1.1	H.323 . . . . .	40
3.1.2	Session Initiation Protocol (SIP) . . . . .	40
3.2	System Overview . . . . .	41
3.3	Call Detail Records . . . . .	44
3.4	Social Call Graph . . . . .	45
3.5	Social Network Features . . . . .	47
3.5.1	Degree . . . . .	47
3.5.2	Call-Rate . . . . .	48
3.5.3	Call Duration . . . . .	49
3.5.4	Eigen Centrality . . . . .	50
3.5.5	Reciprocity . . . . .	51
3.6	Discussion . . . . .	52
<b>4</b>	<b>Caller-REP: Detecting Unwanted Calls Through Caller's Social Strength</b>	<b>53</b>
4.1	Introduction . . . . .	53
4.2	Motivation . . . . .	56
4.3	Caller-REP: Caller Classification-Based on Reputation . . . . .	57
4.3.1	Requirements for Reputation Based SPIT Detection System . . . . .	57
4.3.2	Data Source . . . . .	57
4.3.3	Subscriber Direct Trust . . . . .	58
4.3.4	Reputation of the Subscriber . . . . .	61
4.3.5	Detection of Spammers . . . . .	62
4.3.6	Caller-REP System Components . . . . .	64
4.3.7	Caller-REP and Privacy . . . . .	65
4.3.8	Comparison with Other Reputation-based SPIT Detection Systems	67
4.4	Experimental Methodology . . . . .	69
4.4.1	Synthetic Data-Set . . . . .	69
4.4.2	Evaluation Metrics . . . . .	72
4.5	Performance Evaluation . . . . .	72
4.5.1	True Positive Rate . . . . .	73
4.5.2	False Positive Rate . . . . .	75
4.5.3	Detection Accuracy . . . . .	77
4.5.4	Sparse Subscriber's Network . . . . .	78

4.5.5	Subscriber Reputation . . . . .	79
4.5.6	Caller-REP Vs. Call-Rank . . . . .	80
4.5.7	Caller-REP under Legitimate Network . . . . .	81
4.5.8	Caller-REP under High SPIT Rate . . . . .	82
4.6	Discussion on Caller-REP . . . . .	82
4.6.1	Features of Caller-REP . . . . .	83
4.6.2	Sybil Attack . . . . .	84
4.6.3	Betrayal Subscribers . . . . .	85
4.6.4	Addition of New Subscriber . . . . .	86
4.6.5	Deploying Caller-REP in a Real VoIP Network . . . . .	86
4.6.6	Call Setup Delay . . . . .	87
4.7	Conclusions . . . . .	88
<b>5</b>	<b>COSDS: Blocking Spammers with Information Sharing across Multiple Service Providers</b>	<b>90</b>
5.1	Introduction . . . . .	90
5.2	Limitations of Stand-alone Detection Systems . . . . .	94
5.3	Motivation . . . . .	94
5.4	Collaborative SPIT detection System . . . . .	95
5.4.1	System Design Overview . . . . .	96
5.4.2	Global Reputation of a Subscriber . . . . .	97
5.4.3	Detection of SPIT Subscriber . . . . .	100
5.4.4	Design Options for Information Summarization and Collaboration	102
5.4.5	Communication overhead . . . . .	104
5.4.6	Discussion on Privacy Protection . . . . .	104
5.5	Experimental Methodology . . . . .	107
5.5.1	Synthetic Data-Set . . . . .	107
5.5.2	Evaluation Metrics . . . . .	109
5.6	Performance Evaluation . . . . .	109
5.6.1	True Positive Rate . . . . .	110
5.6.2	False Positive Rate . . . . .	112
5.6.3	Detection Accuracy . . . . .	113
5.6.4	Information Summarization and System Performance . . . . .	115
5.6.5	Effect of Threshold on Performance . . . . .	116
5.6.6	Resilience Against Different Spam Calling Behaviors . . . . .	118
5.6.7	Privacy Breach Analysis . . . . .	121
5.7	Discussion on COSDS System . . . . .	123

5.8	Conclusions . . . . .	123
<b>6</b>	<b>EIS: Early Identification of Spammers</b>	<b>125</b>
6.1	Introduction . . . . .	125
6.2	Motivation and Problem Definition . . . . .	128
6.2.1	Spammer Network . . . . .	128
6.2.2	Why Users Have More Than One Identity . . . . .	129
6.2.3	Motivation . . . . .	130
6.2.4	Problem Definition . . . . .	131
6.2.5	Background Definitions . . . . .	132
6.3	EIS: Early Identification of Spammers through Identity linking and Reputation Aggregation . . . . .	133
6.3.1	ID-CONNECT Module . . . . .	134
6.3.2	Reputation Module . . . . .	139
6.3.3	Spam Detection Module . . . . .	141
6.4	Experimental Data Set and Evaluation Parameters . . . . .	141
6.4.1	Analysis of ID-CONNECT . . . . .	142
6.4.2	Application to Spam Detection . . . . .	149
6.5	Discussion and Limitations of EIS . . . . .	153
6.6	Conclusions . . . . .	156
<b>7</b>	<b>Conclusions</b>	<b>157</b>
7.1	Contributions . . . . .	157
7.2	Future Works . . . . .	160

# List of Figures

1.1	Subscribers Forecast for the VoIP Technology 2013-2018 [INF15]. . . . .	2
1.2	Spammer's Network Model. . . . .	5
2.1	A Taxonomy of SPIT Detection Systems. . . . .	20
3.1	SIP Based VoIP Network. . . . .	40
3.2	SIP Session Establishment and Termination. . . . .	42
3.3	Block Diagram of SPIT Detection System. . . . .	44
3.4	Social Network of Subscribers Extracted from the CDRs. . . . .	45
4.1	Building Blocks of Caller-REP. . . . .	59
4.2	Interaction Between Caller-REP and Proxy-Server. . . . .	66
4.3	Caller Distribution: A) Caller Out-Degree to In-Degree; B) Caller Out-Degree to Out-Duration; C) Caller Reputation to Out-Degree Using Caller-REP. . . . .	71
4.4	True Positive Rate Increases with Time: A) SPIT Rate of 10%; B) SPIT Rate of 20%; C) SPIT Rate of 30%. . . . .	74
4.5	False Positive Rate Decreases with Time: A)SPIT Rate of 10%; B) SPIT Rate of 20%; C) SPIT Rate of 30%. . . . .	75
4.6	Caller-REP Accuracy: A)SPIT Rate of 10%; B) SPIT Rate of 20%; C) SPIT Rate of 30%. . . . .	77
4.7	Caller Reputation With The Time: A) Caller-REP; B) Call-Rank. . . . .	80
4.8	Caller-REP Performance Under Legitimate Network. . . . .	81
4.9	True Positive Rate Under High SPIT Traffic and $\beta=2$ . . . . .	83
5.1	Collaboration Methods: A) Non-Collaboration; B) Distributed Collaboration; C) Centralized Collaboration. . . . .	91
5.2	Building Block of Collaborative SPIT Detection. . . . .	96
5.3	SP's Level Working of Collaborative SPIT Detection. . . . .	97
5.4	Collaborative Simulation Model. . . . .	108

5.5	True Positive Rate of COSDS for SP trust=1 and $\beta$ threshold=1. . . . .	111
5.6	False Positive Rate of COSDS for SP trust=1 and $\beta$ threshold=1. . . . .	113
5.7	Detection Accuracy for COSDS and non-collaborative system for SP trust=1 and $\beta$ threshold=1. . . . .	114
5.8	Information Summarization and True-Positive, False Positive and Accuracy trade-off for different Collaboration Methods. . . . .	116
5.9	The Effect of Threshold $\beta$ for TP and FP Rates for 5 Collaborators and for the First Day. . . . .	117
5.10	System Behavior Against Spammers Having High Out-Degree and High Duration Calls. . . . .	118
5.11	System Behavior Against Spammers Having Small Out-Degree and Small Duration Calls. . . . .	119
5.12	System Behavior against Spammers Having Small Out-Degree and Long Duration Calls. . . . .	120
5.13	Privacy Breach Analysis for Different Scenarios: A) Probability of Breach for Some Auxiliary Information at SP; B) Percentage of Subscribers whose Relationship network identified to some percentage varying number of reputed subscriber. . . . .	121
6.1	Attack Network of Spammer. . . . .	129
6.2	The Calling network of physical individual for two different time periods with two different identities $ID_1$ and $ID_2$ . . . . .	132
6.3	Building Block of EIS System Consisting of Three Major Modules. . . . .	134
6.4	The work-flow of ID-CONNECT for identity linking. The approach outputs list of all identities from $CDR_1$ which are similar to a given identity from $CDR_2$ . . . . .	135
6.5	An example Weighted Call Graph for a given identity ID2 from T2 and two identities (ID1 and ID3) from T1. Without link weights ID2 is similar to ID1 and ID3 but when link weights are considered then ID2 is more similar to ID1 than ID3 because of similar link weights. For ease of reading, rather than showing the $WG_{SR}$ weight directly on each edge, we show vector ( $CallRate_{SR}, \Sigma CDS_{SR}$ ). . . . .	138
6.6	Performance Results for Different Threshold and Barabási-Albert. . . . .	145
6.7	Performance Results for Erdős Rényi. . . . .	146
6.8	Performance Results for Small World Network. . . . .	147
6.9	Performance Results for 80% Mutual Friends. . . . .	148
6.10	Performance Results for 50% Mutual Friends. . . . .	148

6.11 Performance Results for 30% Mutual Friends. . . . .	148
6.12 False Positive Rate for A) Barabási-Albert, B) Erdös Réenyi, and 3) Small World Networks. . . . .	150
6.13 True Positive Rate for Different Overlaps in Victim Network. . . . .	152
6.14 False Positive Rate for Different Overlaps in Victim Network. . . . .	154
6.15 Accuracy for Different Overlaps in Victim Network. . . . .	155

# List of Tables

4.1	Confusion Matrix. . . . .	72
5.1	Subscriber Level Privacy Breach for Different Collaboration Methods. . .	106
5.2	Service Provider's Level Privacy Breach for Different Collaboration Methods. . . . .	106

# Abbreviations and Symbols

## Abbreviations

ACC	Accuracy
Aux	Auxiliary Information
BA	Barabási-Albert
C/R	Challenge/Response
CallerREP	Caller Reputation
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CDR	Call Detail Records
CL	Candidate List
CON	Out Mutual Friends Connections
COSDS	Collaborative SPIT Detection System
CR	Centralized Repository
CSS	Candidate Set Size
DDoS	Distributed Denial of Service Attack
DoS	Denial of Service Attack
DTMF	Dual Tone - Multi Frequency
EC	Eigen Centrality
EIS	Early Identification of Spammers
ER	Erdős Réenyi
FN	False Negative
FP	False Positive
FTC	Federal Trade Commission
HTML	HyperText Markup Language
ID	Identity
IDC	ID-CONNECT
IETF	Internet Engineering Task Force
IL	Identity Linking
IM	Instant Messaging

IP	Internet Protocol
ITU	International Telecommunication Union
MCU	Multipoint Control Units
MF	Mutual Friends
MFS	Mutual Friend Similarity
MGCP	Media Gateway Control Protocol
NGN	Next Generation Network
OMF	Out Mutual Friends
OSR	Out Sim-Rank
PR	Privacy
PSTN	Public Switched Telephone Network
SAS	Stand-Alone System
SCCP	Skinny Client Protocol
SIP	Session Initiation Protocol
SMC	Multi-party Computation
SMS	Short Message Service
SP	Service Provider
SPIT	Spam over Internet Telephony
TN	True Negative
TP	True Positive
RTP	Real-time Transport Protocol
RTCP	Real-time Transport Control Protocol
UA	User Agent
USA	United States of America
VoIP	Voice over IP
WG	Weighted Call Graph
WS	Weighted Similarity
WS	Watts-Strogatz

### **Symbols**

A	Total Number of Subscribers.
S	Caller who initiates the call request.
R	Callee who receive the call request.
G	Call Graph G between caller and the callee.
DT	Direct Trust
LR	Local Reputation
GR	Global Reputation

m	25th percentile of reputation scores.
T	Time period.
$\beta$	Parameter defined by SP.
CR	Centralized Repository
OD	Out-Degree
CD	Call Duration
$P_i$	Individual who has multiple identities.

# Chapter 1

## Introduction

Voice over IP (VoIP) - an Internet Protocol (IP)-based voice communication system is increasingly used by a large number of people along with a traditional circuit switched network (mobile, landline) for business and personal communications. In recent years, VoIP has seen an enormous growth in the number of subscribers because it offers affordable calling rates for any destination across the world. Moreover, it provides affordable value services and flexibility of using IP networks for the voice communication. VoIP market is expected to reach more than 1200 million subscribers worldwide by 2018 [CIS18] with the expected revenue of more than \$77 billion [INF15]. Figure 1.1 depicts the growth of residential VoIP subscribers from year 2013 to year 2018. The number of business subscribers is also increasing at the rate of 7.58% and would reach to 244 million business subscribers by 2018 [CIS18]. The affordable calling rates of VoIP, its easy integration with the IP networks, and value added services has also created a lucrative opportunity for spammers and telemarketers to make the unwanted, bulk un-solicited calls via VoIP. In VoIP these calls are referred as SPIT (SPam over Internet Telephony (SPIT) and are mainly used for advertising products, harassing subscribers, convincing subscribers to dial premium numbers, making *Vishing*(voice equivalent of web Phishing) attack to get private information of call recipients etc. Spammers can also make unwanted calls to steal user's information [NNS+07], make calls to check unsecure gateways within the service provider for the termination of bulk un-billed calls [Zwy+07], and cause disruption in the network services through flooding and denial of service attacks [EGM10, KER11].

Unwanted phone calls and instant text messages can come at any hour of the day. These unwanted calls and instant messages require immediate response from the recipient, thus annoy call recipients while at work, disturb them in their family times, and can even interrupt sleep in late hours at night. Recent statistics on telephony spam have revealed that answering a spam call would result in an estimated loss of 20 million man

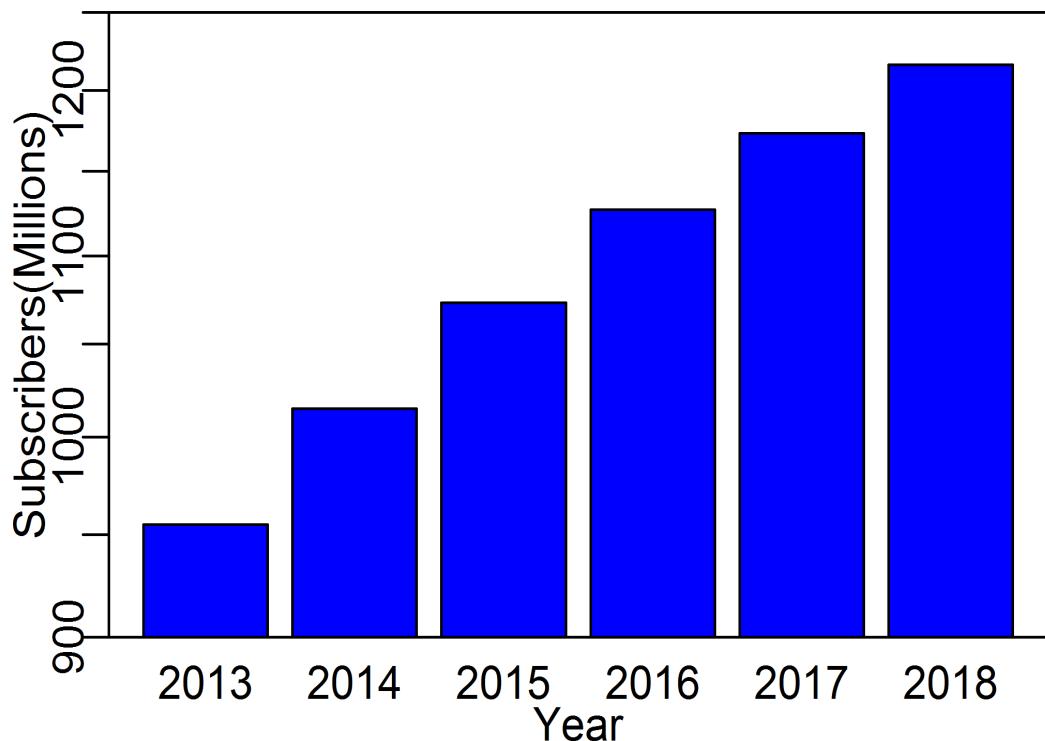


Figure 1.1: Subscribers Forecast for the VoIP Technology 2013-2018 [INF15].

hours for a small business enterprise in the United States with the estimated loss of about \$475 million annually [SPA2015]. Moreover, FTC (Federal Trade Communication) has estimated that every year scammers and spammers causes a loss of \$8.6 billion annually to citizen of USA due to frauds and majority of them are initiated from the telephone. Every year service providers, regulators, and law enforcement agencies receive thousands of complaints from consumers for unsolicited, unauthorized, and fraudulent callers trying to abuse them. In 2012, FTC (Federal Trade Communication) in USA has received four times more complaints against unwanted calls than number of complaints they received in 2010 [JEN15]. The number of identified spam callers has also risen to 162% from January 2013 to January 2014 and call center fraud has risen to 45% since 2013 [PIN16].

## 1.1 Spam Over Internet Telephony

Voice spam or SPIT (Spam over Internet Telephony) are the unwanted, unsolicited, pre-recorded advertisement phone calls made by the spam sender to a large number of recipients that has no prior social relationship with the recipients. VoIP spammers are similar

to email spammers as both have the same intent of delivering unsolicited information to the recipients that contains advertisements of legal or illegal products, commit fraud to recipients by getting their private information, and spread viruses. The spam calls and messages can also be sent to and from the mobile telephony system and the legacy PSTN (Public Switched Telephone Network) telephony. The following are additional forms of spam introduced because of telephony:

**Instant Message Spam:** Bulk unsolicited instant messages (similar to email spam messages) but sent instantly to users of messaging system like Skype [RSM06], WhatsApp, and Viber etc.

**Presence Spam:** Presence spam is a bulk unsolicited set of presence requests messages sent to the subscribers in order to get subscribers buddy or white list for sending IM and call spam.

**Virus Spam:** Sending viruses inside bulk SMS or IM messages that affects operating system of phones and discloses system vulnerabilities to spammers.

### 1.1.1 Consequences of Spam over Internet Telephony

SPIT is one of the interactive forms of network abuse, where call recipient is required to respond immediately for the incoming call. Unlike email spammer, VoIP spammers are not only irritating to recipients but also would cause significant loss because of answering a spam call while roaming and use of other value added services while answering the spam call. In telephony, spammers can be a threat to the subscribers of technology for the following reasons [3GPP2015]:

**Callee Account Credit** Telephony subscribers pay extra amount for the value added services such as call forwarding, roaming, automatic call back, etc. The receiving of unsolicited calls and messages while roaming would charge call recipients for the nothing. Similarly, automatic call back service might result in a call back to some premium numbers without user's intentions.

**Missing Important Calls** Legitimate users normally do not want to miss important calls and forward them to the voice mail box during their periods of unavailability. However, a typical voice mail box has very limited storage capacity which can be overwhelmed by the spam call in the form of recorded message. This would result in resource unavailability for the calls from legitimate callers being forwarded to the voice mail box. It would also be irritating and a waste of time for the recipient to

go through each spam message recorded in a voice mail box. More importantly, recipient would also likely to miss some important recorded message if he dismissed them as spam on fly or if the mail box already full with the spam messages.

**Vishing** *Vishing* is equivalent to Phishing in the Web. In *Vishing* attack, spammer tries to steal someone's private information by impersonating as a legitimate entity and later use this information for the fraud activity. In email network, Phishing might have small impact because recipients have some time to consult others before responding to the message, but in telephony recipients need to decide immediately otherwise they would miss the opportunity of getting some financial benefits.

**Financial Loss to Service Provider** Spam traffic consumes bandwidth and network resources, making resources unavailable to the legitimate users. Additionally, spammers or fraudsters try to identify some open and non-secure service providers for un-billed call termination. This would result in a serious financial damage to service provider if spammer remains undetected for a long time periods. In the perspective of service provider's Quality of Service, users become annoyed if they do not receive sufficient network resources at the time of their calls. They also become annoyed if they receive large number unwanted advertisement calls without their consent. These scenarios would increase distrust of users on their service provider and continuous bad service would convenience them to change their service provider.

### 1.1.2 Motivation of VoIP Spammer

The major motivation for all spammers is get financial benefits by making frauds with the recipients of calls and messages. Spammers make spamming attempts for advertising legal and illegal products, disseminating religious and political campaigns, convincing recipients to buy product, call back to the premium numbers or disclose their private information for prize etc. The main goal of the spammers is to deliver their message to a large number of people for greater financial benefits with small investment. Figure 1.2 depicts the spamming model of the spam caller in a VoIP network. A SPIT caller can be categorized into two groups: 1) auto dialer – where an automated machine or a computer generates a large number of spam calls to a large number of recipients, and 2) a human SPIT caller – where the spammers hire a cheap labor for making unsolicited calls to a large number of recipients. In terms of financial benefits, spammers gain benefits in three ways. Firstly, they convince the callee to call them back on the premium numbers. For this purpose, spammers scam people by using social engineering attacks and exploit the needs of people in the particular societies such as offering expensive gifts and attractive tourism

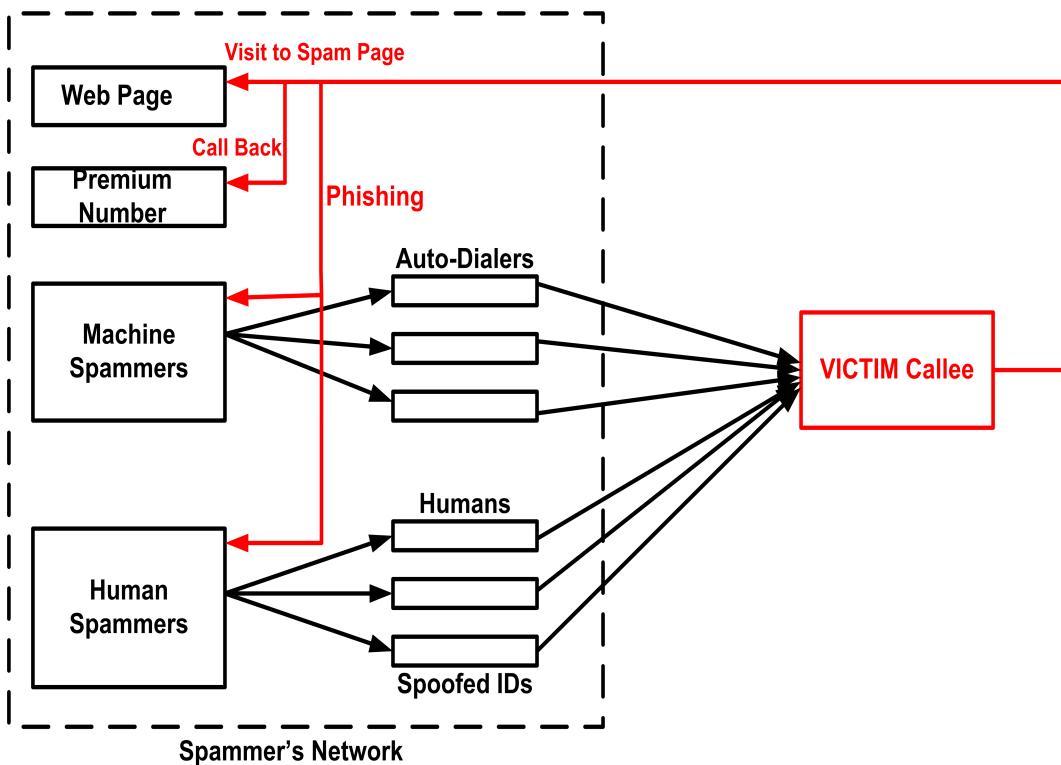


Figure 1.2: Spammer's Network Model.

packages to historical and religious places. Secondly, they convince the callee to disclose his private information by impersonating as a legitimate entity. Thirdly, they make callee to listen to the complete advertisement and convenience them to buy products.

Spammers can also make spam calls or send spam messages for scanning end-user handset for the system vulnerabilities. The ability of making anonymous calls or miscalls to the specific callees from VoIP network also encourages spammers to use VoIP medium for threatening or annoying recipients with the denial of service attack to a specific recipients. Beside financial benefits, spammers can also use telephony to spread a real time interactive religious hate or political messages to a large number of users. In perspective of service provider or organization, spammers can make spam calls to learn the system vulnerabilities and identify some open gateways for a free of cost call terminations.

### 1.1.3 SPIT Differences from E-mail Spam

SPIT exhibits some similarity with the email spam. Both email spammer and a SPIT caller uses Internet as a medium for conveying his core message to recipients but SPIT causes more serious discomfort to the victims because of the real time response for the call. In telephony, the call recipient has to decide immediately whether to accept or ignore

the call. A callee has already been affected (waste of time) after realizing that the received call is a spam call. Besides similarity in motivations and IP medium, SPIT exhibits some differences from the email spams. E-mail spammer utilizes text messages, images or attachments for conveying their message to the victims, whereas SPIT caller uses digitized speech streams over the Internet for conveying his messages. In terms of deciding about sender and contents inside the message, the email service provider can hold e-mails for a sometime before finally delivering them to the recipient's inbox, which is not noticeable to recipient. The VoIP or Voice service provider cannot hold speech streams and signaling messages without addition of noticeable delay in a signaling and flow of speech streams between users. In perspective of content processing, online processing of speech content is more challenging and resource intensive than offline processing of text messages and images.

From the user's perspective, a single e-mail spam can remain in the inbox unattended for as much time as the user wishes, but in the case of SPIT or voice call user has to respond back interactively, which makes it more annoying and disturbing. In the perspective of user's resources, a single spam email typically consumes small number of bytes, but a voice message in a voice mail box requires much greater space thus making voice mail box unavailable for the legitimate calls.

In terms of human effort, the deletion of a SPIT call is more annoying and intrusive than the deletion of spam emails. In email network, firstly, the service provider assists end-user in classifying senders, and secondly end-user decides about email on a first look by reading the subject. On the other hand, in telephony, a user is required to listen the recorded call before thrashing it away as spam. The detection and deletion of spam speech content from the voice mail box is more time consuming as it requires at least 6 steps to completely remove the speech content from the voice mail box [TDZ+16]. Moreover, there exists no system that allows service providers to inform callee about the nature of calls recorded in the voice mail box. Additionally, in telephony user might also delete some important calls if he is making decision on a fly. In a perspective of protocol architecture, an E-mail message is composed of two parts: a header and a body. The email header part can also provide some information about sender's nature. Telephony calls also consists of two parts: a call setup phase and a speech streaming phase; but the messages exchange in call setup phase though are available in a plain text but are not providing any information about the sender's nature and the speech stream is only available after the call setup.

### 1.1.4 Why is it Hard to Detect SPIT Caller ?

The affordable calling rates and use of telephony services over the Internet has convinced many subscribers and organizations to adopt VoIP as media for the business and non-business communications. Unfortunately, these features of VoIP have also attracted spammers to make use of this media for unsolicited activities to a large number of victims interactively. To make VoIP usable, to improve productive and trustworthiness of services, it is extremely important for the service provider to identifies and blocks spammers in a timely way. Unlike other forms of spams (for example email spam, blog spam, web spam) VoIP spam or SIPT is much more difficult to detect. This is because spam is spread with the use of speech streams, which is only available after the call setup. Moreover, the service provider typically does not allocate sophisticated network resources for the processing of large number of speech samples in a real-time. Though a human user can distinguish spam speech from non-speech, but this is always late as spammer has already annoyed recipients with the unwanted content. The design of an effective SPIT detection system for a VoIP service provider is a challenging task because of the following reasons:

**Non-Availability of Speech Contents Before Call-Setup:** A typical VoIP call consist of two phases: 1) a call setup phase where a call request messages are exchange between caller, callee and the service provider, and 2) a real-time media exchange phase where a speech stream is exchange after a successful completion of call setup process. The messages that exchanged between caller, callee and proxy server during the call setup phase are in plain text and present call handling properties of involved parties. Though the signaling messages contain some information about caller and the callee, but they cannot provide enough information to be used for identifying SPIT caller. Additionally, it is very easy for the spammer to change his signaling messages and make them similar to the signaling messages of the legitimate users, thus leaving signaling content impractical for the spam detection. The speech contents could provide information about whether caller is spammer or not but speech content is only available after the call setup hence very late as spammer has already annoyed the callee with spam content. Beside this limitation, speech content- based spam detection systems have other limitations. Firstly, processing real-time speech content requires sophisticated system resources for a real-time processing of speech content. Secondly, speech processing on active stream flows would add unnecessary delays to the conversations. Thirdly, processing content is prohibited by law in many countries. In spammer's perspective, content-based approaches can be easily evaded by the spammers by adding a random noise in the speech streams.

**Intrusiveness:** Callers and the callees become annoyed and irritated if they are repeatedly asked for solving a certain challenge or asked to provide feedback about the

caller's at the end of every call transactions. Existing SPIT detection approaches involve callee in two ways: 1) after the call termination, and 2) before the call establishment. In the first way, callees are being asked to provide positive and negative feedback [KD07], [DK05], [WBS+09] about the caller's transaction as soon as call ends. This feedback is then used for computing global reputation of the caller within the network. In the second way, the service provider provides credentials of the caller to the callee who can then decides whether to accept or reject the call [BAP07]. Both of these approaches are intrusive and require changes in a VoIP infrastructure: for example the addition of email like spam button in the VoIP or telephony handset. In terms of caller, the existing approaches ask caller to prove authentication in two ways. First, the caller is asked to solve a certain challenge in the form of CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) or Turing test, and second, the caller is asked to authenticate himself through the exchange of valid private and public keys. The first approach is intrusive while the second approach though is non-intrusive but require private-public infrastructure for the authentication and authorization. Moreover, the challenge-response-based approaches require additional system resources for handling a large number of concurrent callers and would significantly add call setup delay.

**Calling Behavior and Automatic Classification:** Telephony has become a popular medium for the communication. Through Telephony, users talk to their family and friends, and conduct businesses. A clear understanding of the communication behavior of spammers and non-spammers can help for the effective design of a SPIT detection system for the VoIP network. Calling behavior of users such as call-rates, the number of unique callees of the user, the time duration of user's interactions with their peer callees, and the number and duration of incoming calls to the user can significantly help in modeling the behavior of user. The use of single feature for modeling behavior of user would not provide an effective resistance against spammers, as spammers can easily evade them by manipulating the single feature. For example, SPIT detection approaches proposed in [BAP07], [RSM06] use average call duration for computing global reputation of the user for classifying user as a spammer and a non-spammer. However, these systems can be evaded by spammers through the creation of a Sybil network among their own identities. Moreover, the use of an average call duration feature would assign high reputation scores to the caller having small duration calls with majority of his callees despite having large number of recipients and non-repetitive calls. The challenge in an effective behavioral-based spam detection is three-fold: first investigating which set of features are difficult to be manipulated by spammers, second, investigating which features could be used collectively for the computation of global reputation, and third having the system that automatically decides about nature of the user without any user intervention.

**Privacy-Aware Collaboration:** Existing standalone SPIT detection systems consider locally recorded call logs for modeling the behavior of the users within the service provider network. The standalone detection systems lack global view of user's behavior in other service providers or home service provider. These systems can prolong the detection of spammers that make low rate spam calls to the recipients of particular service provider, but distribute calls to the recipients of many service providers. Collaboration among service providers would naturally improve the detection time and detection accuracy, but this depends on the amount of information exchanged during the collaboration process. Service providers are reluctant in talking part in collaborating with other service providers because they are business competitors and are concerned about privacy of their customers and their internal network configuration. The challenge in the design of an effective collaborative SPIT detection systems is three-fold: first, determine what filtered information should be exchanged among collaborators so that service provider remains confident and take parts in the collaboration, second, understand with whom this information should be exchanged such that privacy of collaborating service provider is not compromised, and third, determine what information should be return back to the collaborating service providers.

## 1.2 Behavior-based Collaborative SPIT Detection System

There exist several reputation-based detection systems that use identity of a subscriber<sup>1</sup> for computing reputation of the subscriber within the service provider network. The core concept in the reputation-based systems is simple that is to use the behavioral communication patterns of the subscriber and classifies subscriber as a spammer if subscriber has abnormal communication patterns. These approaches are based on the fact that spammers and non-spammers exhibit different calling behavior towards their friends and non-friends. The reputation-based approaches have shown some effectiveness in filtering spammers in an email and the social networks [KRS+06], [LY07]. These approaches mainly considered social network features such as total number of recipients, the clustering coefficient, reciprocity and bi-directional communication patterns of subscriber in a network. However, in a voice network, call duration and call-rate are additional features that can provide additional information about the relationship strength between subscribers. In a telephony, legitimate subscriber normally exhibit repetitive calling behavior with their family members and friends with a relative long duration calls [BSG+11]. The

---

<sup>1</sup>subscriber, individual, users terms are interchangeable in this thesis. A subscriber can be a caller who initiates a call, can be callee who receive a call or both.

legitimate subscribers also receive repetitive long duration calls from their family members and friends. However, on the other hand spammers normally try to have a larger footprint by targeting large number of users that result in a large number of small duration calls and non-repetitive calling behavior. This behavior is due to the fact that spammers crawls the web or a telephone directory for their victims or randomly generate large of identities of a specific series and large number of victims are not interested in talking for a long time.

In a VoIP network, the performance of behavioral-based anti-SPIT systems depend on the type of social network features used for modeling the behavior and computing reputation of the subscribers. The challenge in reputation-based approaches is to ensure that they do not involve subscriber at any stage of call processing. In terms of selection of social network features these approaches require to use number of social network features collectively rather than using a single social network feature alone. Several reputation-based approaches have been proposed for filtering SPIT callers in a VoIP network [KD07], [DK05], [WBS+09], [BAP07], [RSM06], [RS05]. These approaches compute reputation of subscribers in two steps. First, a direct trust between subscriber and his called subscribers is computed from the subscriber's past call transactions with others. Second, a global reputation of the subscriber is computed by aggregating the direct trust scores of a subscriber. The direct trust scores between subscribers represents a level of direct relationship between them and can be computed explicitly by getting feedback from the called subscriber after the end of call transaction [KD07], [DK05], [WBS+09], or implicitly use information from the call logs recorded in the call detail records [BAP07], [BSG+11]. The global reputation represents aggregate calling behavior of the subscriber by considering the behavior of subscriber towards all interacted subscribers.

Presently, the existing reputation-based SPIT filters are intrusive to subscribers which not only annoy subscribers but also required changes in a VoIP handset and signaling messages. A non-intrusive reputation-based SPIT detection system computes reputation of subscriber by extracting information from the user's call logs for his past transactions. Call-Rank [BAP07] is a reputation-based spit detection system that computes global reputation of the subscriber by leveraging average call duration feature between subscriber and his callees, and the Eigen trust algorithm. Additionally, Call-Rank asks callee for the final decision about whether to accept a call or reject a call by relaying caller's reputation scores and caller's social network credentials to the callees. This approach has few limitations. First, it only leverages average call duration feature for computing the direct trust that can be prone to be evaded by the spammers targeting large number of callees and managing good duration calls with only few of them. Secondly, Call-Rank is intrusive to the caller during the stage of CAPTCHA test and intrusive to the callee during the deci-

sion phase. Third, it discloses social network credentials of caller to the callee that can breach privacy of the caller. Fourth, it requires public and private key infrastructure for authentication which is difficult to be implemented in VoIP telephony. In [ZG09], authors used call duration with a threshold of 20 second as sign whether caller is spamming or not. The system considered caller call as reputed if call duration of the call is greater than 20 seconds and consider caller call as a non-reputed if call duration of the call is less than 20 seconds. However, we believe spammers would easily evade such system and would also have high false positive if legitimate callers also have calls less than 20 seconds. In terms of feature usage, the existing behavior-based approaches use only one feature for the computing reputation of the caller. However, we believe that collective use of features for computing reputation of the caller would significantly improve the effectiveness of a SPIT detection system without involving caller and the callee. Our objectives for the design of reputation-based SPIT filtering system for the VoIP and voice network is three-folds: 1) it must be non-intrusive (does not require any interaction with caller and callee at any stage of call processing); 2) it collectively uses call duration, call rate, and total number of recipients of the caller for computing direct trust and reputation of the caller, and 3) it automatically learns the classification threshold below which the caller is considered as a spammer. To achieve the objective of non-intrusive, feature rich collaborative behavioral-based SPIT filtering system we argue the following:

1. Direct trust and global reputation of caller: Subscribers usually develop two types of connections over the time: strong connections and weak connections. Strong connections are established with friends, family members and colleagues with whom subscriber interacts more frequently and for a long time duration. Weak connections are established with the people with whom subscriber interacts less frequently and with the smaller call durations. Call duration and call-rate between subscriber and his friends are important features that can provide information about the strength of the trust between subscriber and his called friends. However, we believe it should not be limited to call duration and call-rate in one direction (caller to callee) but should also incorporate call duration and interaction rate in both directions. A fundamental difference in call behavior of spammers and non-spammers is that spammers normally target an exceptionally large number of callees, whereas non-spammers have limited number of called callees. The use of total number of recipients would discriminate spammer from non-spammer but using it alone might be prone to be bypassed by the spammer and also have false positives. We believe that the collective use of out-degree, call rate, and call duration in both directions for computing direct trust and global reputation of a caller would probably result in a

small direct trust score and small global reputation score to spammers and a high reputation scores for the legitimate users.

2. Decision and Threshold Selection: The legitimate subscribers typically have high global reputation scores because of their strong connections with their callees. On the other hand, spammers have small reputation scores due to imbalanced in their communications i.e. large number of recipients, small incoming calls and small duration calls. Existing SPIT detection approaches rely on the callee for making a final decision about acceptance or rejection of the call, which is intrusive and require changes in a signaling messages. To reduce the interaction with the callee, the detection and reaction system within the service provider need to decide automatically whether to allow or block the caller using automated threshold. Furthermore, the threshold should be tunable and scalable so that it can be easily integrated with other social network features for the improved detection accuracy.
3. Service Provider Collaboration: Intelligent and smart spammers disperse their spam calls across many service providers without overwhelming a single service provider. These small rate spammers remain undetected by the standalone SPIT detection systems for a relative long time period. However, the behavior of spammers normally remains same across all attacked service providers, thus having collaboration among service providers would possibly improve the detection time and accuracy. However, convincing service providers to collaborate is a challenging task because service providers are reluctant in sharing information that may affect their network and customer privacy. There is a strong need for a resource intensive collaborative system that convinces service providers to be part of collaboration without worrying about the privacy of their customers.
4. Identity-Linking: A high number of spammers only make few spam calls to the recipients. This is because of two facts: first the spammer acquired a large number of identities and is targeting users from his different identities; and secondly, service providers block spammers for his spamming activity. Moreover, once service provider blocks certain spammer, the spammer either whitewashes his reputation scores or starts making spamming from a new identity. Though spammers change their identities but their target remains the same and there is a possibility that spammer has overlap in targets from their different identities. The linking of different identities of spammers would help in early detection of spammers that make small rate intelligent spamming to a large number of recipients from their different identities over time. Moreover, identity linking would also help in characterizing social

behavior of people with more than one identity and possibly identifies criminal's rings.

## 1.3 Thesis Contributions

The thesis has three major contributions. First, we address the problem of spam from the perspective of standalone service provider, second, we involve service providers to take part in collaboration for early detection of spammer, and third we propose a method for identity linking in a voice networks. The major contributions of this thesis can be summarized as follows:

**Standalone SPIT Detection System:** A non-intrusive reputation-based standalone SPIT detection system has been proposed that uses past calling behavior of subscriber modeled from the call details record of subscriber. To achieve objectives of an effective and non-intrusive standalone SPIT detection system, we contribute the following:

- We present a method for computing direct trust between caller and the callee, and a method computing global reputation of the caller in a service provider. The proposed method explicitly uses information from the subscriber's past call transactions and computes global reputation of the caller in two steps. In the first step, a direct trust between caller and the callee is computed by using three features collectively, namely: call-rate in both directions (caller to callee and callee to caller), call duration in the both directions and total number of callees of the caller. In a second step, a global reputation of the caller is computed by aggregating the normalized direct trust score of caller with his callee using modified power iteration algorithm. The computation of reputation in this way would assign a high global reputation scores to the legitimate callers and a small reputation scores to the spammers because of their non-connected social network and large number of target callees.
- We present an automatic procedure for computing dynamic threshold below which callers are considered as spammers. For the automatic threshold, we adopted a dynamic 25th percentile based threshold that is being computed for each global aggregation cycle. The threshold is scalable and tuneable according to the requirement of the service provider detection policies and can be easily integrated with other call and social network features.

- A model for privacy protection of caller and the callee while computing reputation of the caller from the CDRs (Call Detail Records). CDR contains sensitive information about the social relationships network of subscriber with the other and the subscriber requires absolute protection of his data and relationship information. The privacy of the subscriber is protected by sending filtered information to the reputation engine. To achieve this objective, the proposed system adopts the following: first, the task of reputation computation and the detection is carried out on an independent system separated from the proxy server or main CDRs database, and second, a filtered and anonymized CDR is exchanged with the reputation and detection system. The detection and reputation engine responds back with the final results about subscriber without disclosing his friendship network.
- Telecommunication data-sets are not available to analyze the performance of a detection system. In order to evaluate the performance of propose system, a detailed synthetic model is therefore required for the generation of synthetic data-set that can characterize the behavior of spammers and non-spammers in a real telecommunication network. To achieve this objective, we developed a comprehensive synthetic model for the generation of synthetic data-set that considers social behavior of spammers and non-spammers in terms of call-rates, call durations, and out-degree distribution. The synthetic data-set is generated for a number of days and for different percentages of spammers and non-spammers. Finally, the standalone SPIT detection system is evaluated for the different performance metrics that are accuracy, true positive rate, false positive rate etc.

**Collaborative SPIT detection System:** A collaborative SPIT detection system is presented that aggregates information from the collaborating service providers without compromising privacy of collaborators and their customers. To achieve objectives of a privacy-aware collaborative SPIT detection system we contribute with the following:

- A privacy aware collaborative model is presented for the exchange of information among collaborating service providers. The proposed model protects privacy of the collaborators and their customers through the use of trusted centralized repository and through the exchange of non-sensitive information to the centralized repository. The centralized repository computes global reputation of the subscriber by applying a weighted average algorithm and updates

collaborating service providers with the reputation scores and classification decisions. The intruder or an adversary at a centralized repository would not be able to infer social relationship network of subscribers and would also not be able to learn private information of collaborating service providers.

- A procedure for aggregation of received reputation scores from the collaborating service providers and a detection decision. The proposed procedure uses a weighted averaging mechanism for the reputation aggregation. Moreover, a procedure for the computation of trust among service providers has also been proposed. We have also analyzed the privacy breach analysis for different adversarial information.
- A procedure for the generation of extended synthetic data-set which considers users behavior within their home network and the visiting networks. The model is an extension of the model presented in contribution 1 and incorporates calls among different service providers. The collaborative SPIT detection system has also been analyzed for different numbers of collaborators and for different calling behavior of spammers.

**Identity Linking and Early Identification of Spammer:** An Identity linking system that connects multiple identities of a single individual within the service provider. In developing this, we make the following contributions:

- We introduce an ID-CONNECT system, a social network and behavior based model that links similar identities that probably belongs to a one physical individual. In ID-CONNECT, the weights on the links between identities are computed from the interaction rates and length of interactions. Individuals, especially spammers normally exhibit similar call behavior and have overlap in victims from their many identities. Two identities can only be considered as similar if they have common friends have similar calling behavior towards common friends. ID-Connect is a two-step approach: firstly, it estimates the weighted similarity measure between identities by considering the call behavior of identities towards their common friends. Secondly, it generates candidate set for the given identity using fixed thresholds.
- A reputation engine that computes reputation of the individual by using call-rate, call duration and out-degree of the individual after connecting his identities. The input to the reputation engine is the data of linked identities formulated from the ID-CONNECT module. We believe that reputation computed

after linking identities of an individual and analyzing his aggregate calling behavior would greatly separate spammers from the non-spammers.

- A detection module for computing automated classification threshold below which individuals are flagged as spammers. A dynamic automated threshold is being computed for each reputation cycle using the percentile based approach.
- We validate and evaluate EIS system through a comprehensive simulation study using a synthetic data set that we have generated using true behavior of spammers and non-spammers. The experimental results show that EIS system outperforms other identity linking systems and has shown effective resistance against spammers having many identities.

The work presented in this thesis is intended to identify and block spammers in voice and VoIP service providers. The solution is non-intrusive and can be deployed within the service provider as a standalone system without making any substantial changes in the network. Additionally, the standalone system can also be integrated with other solutions, for example Turing or CAPTCHA test as a solution for the improved detection accuracy.

## 1.4 Publications

The results of this thesis have been presented in the following publications:

1. Early Identification of Spammers Through Identity Linking, Social Network and Call Features [**Muhammad Ajmal Azad, Ricardo Morla**] Submitted to Elsevier Journal of Computational Sciences (Major Revision)
2. Blocking Spammers with Information Sharing across Multiple Service Providers [**Muhammad Ajmal Azad, Ricardo Morla**] Submitted to IEEE Transactions on Dependable and Secure Computing (Major Revision).
3. System and Methods for Detecting Spammers in a VoIP Network [**Muhammad Ajmal Azad, Ricardo Morla**] To be Submitted.
4. Caller-REP: Detecting unwanted Calls with Caller Social Strength [**Muhammad Ajmal Azad, Ricardo Morla**] In Elsevier Computers & Security, Volume 39,Part B, pp-219-236.
5. ID-CONNECT: Combining Network and Call Features to Link Different Identities of a User [**Muhammad Ajmal Azad, Ricardo Morla**] In The 18th IEEE Conference on Computational Science and Engineering (18th IEEE CSE, Privacy, Trust and Security Track), October 2015.

6. COSDS: Privacy Aware Collaborative SPIT detection System [**Muhammad Ajmal Azad, Ricardo Morla**] in Ist Symposium on electrical and computer engineering June 2015.
7. Mitigating SPIT with Social Strength [**Muhammad Ajmal Azad, Ricardo Morla**] In The 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) 25-27 June 2012.
8. Multistage SPIT detection in Transit VoIP [**Muhammad Ajmal Azad, Ricardo Morla**] In The 19th International Conference on Software, Telecommunications and Computer Networks (SoftCOM) 14-17 September 2011.

## 1.5 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 describes the state-of-the-art research on the spam detection in a VoIP network. We provide state-of-the-art for the following: Standalone SPIT detection systems, collaborative SPIT detection systems, and identity linking systems. Moreover chapter 2 also discusses advantages and disadvantages of existing systems and motivates for the design of our behavioral-based collaborative SPIT detection system.

Chapter 3 describes social behavior of the spammers and the non-spammers. This chapter also describes call detailed records used for the generation of call graph of the subscribers and examines the behavioral features that better characterize and differentiate spammers from the non-spammers.

Chapter 4 motivates and describes our standalone SPIT detection system. A standalone method is proposed for computing reputation of the subscriber through his call and social network features, and classification threshold is computed below which a subscriber is classified as a spammer.

Chapter 5 motivates the objectives and needs for the collaborative SPIT detection system. A collaborative SPIT detection system is proposed that uses trusted centralized repository and filtered information for the collaboration process.

Chapter 6 presents a mechanism for connecting identities that belong to one physical subscriber. A method is proposed that estimates similarity between two identities from CDRs at two different time periods. Furthermore, the chapter discusses the effect of identity linking on the SPIT detection.

Chapter 7 discusses the impact of this thesis and highlights some future research issues related to security and privacy in a VoIP and voice network.

## **Chapter 2**

# **Systems for Detecting Unwanted Communications in a VoIP Network**

In the past few years, social networks and telephony (mobile, fixed and VoIP) have become the most important form of communication media for instant messaging and interactive communications. Spam over the Internet has long been the problem in the form of email spam that results in an overall loss of tens of billions of Dollars annually. However, recently email spamming has dropped drastically [REP15] as spammers are finding new ways to target users of other technologies such as telephony and social networks with the unsolicited communications. Spamming in an interactive media such as VoIP, instant messaging and traditional circuit switched network is more annoying than email spamming as recipients are required to respond the incoming call request immediately. These unwanted calls not only effect productivity of subscribers but also causes a financial loss to the target victim. Besides gaining financial benefits, spammers also try to distribute malware to infect recipient's mobile and VoIP handsets or to find system vulnerabilities.

In recent years, VoIP telephony has shown tremendous increase in the number of subscribers because of affordable telephony rates and flexible use of Internet technology for a voice communication. It is of utmost importance for the service providers to have an effective SPIT detection system that can significantly contribute telephony subscribers from abuses and frauds. The effective detection of spammers would not only increase productivity of telephony subscribers but also improve trust of subscribers on their service providers.

Several approaches have been proposed for combating spammers in a VoIP network. These approaches can be grouped into two categories: content-based SPIT detection systems and the identity-based SPIT detection systems. The content-based detection systems process the speech streams that are being exchanged between sender and the recipient for

the identification of a famous spam phrase or word. The identity-based SPIT detection systems use identity of the subscriber (calling identity or IP-address) to monitor the behavior of the subscriber within the service provider network. This chapter provides an overview of the prior works that have been carried out for identifying and blocking spammers in a VoIP network. Specifically, this chapter is organized in three major sections: 1) it provides detail discussion on approaches that have been proposed for blocking spammers in a service provider network; 2) it provides detail discussion on the works that have been carried out for collaboration among service providers; and 3) the works that have been performed for linking identities that belong to the single individual across the network.

## 2.1 Stand-alone SPIT detection

The stand-alone spam detection system monitors behavior of users within the jurisdiction of a single service provider. This section outlines some of the standalone SPIT detection approaches that have been proposed for mitigating spammers in a standalone VoIP network. The taxonomy of standalone SPIT detection systems is shown in Figure 2.1, and each block of the Figure 2.1 is described in sections below.

### 2.1.1 Content-based Detection Systems

Content processing has been widely used for mitigation of email spams [UT08], [SV99], detection of spammers and spam content in the blogs [KJF+06], identification of malicious and spam web pages [NNM+06], filtering of spam messages in the online social networks [SML+14], and filtering of SMS spam in a mobile network. The content-based approaches apply sophisticated machine learning mechanisms [UT08], [SV99] to the labelled (spam and non-spam) contents and classify new content as a spam and a non-spam. The content-based approaches have also been used in the VoIP networks for processing speech content in order to find famous spamming phrases and words.

In a VoIP network, content-based approaches can be grouped into two categories: online speech processing and offline speech processing. In online speech processing, the detection system processes the speech content in real-time that is flowing between sender and the recipient, whereas the offline approaches process the speech content stored on the voice mail box and the media servers. Spammers normally replicate their spamming contents to all their target subscribers, which can be stored in the voice mail box of the target subscribers. A content independent offline speech processing method has been proposed in [ISW13] that estimates similarity between speech samples left on the

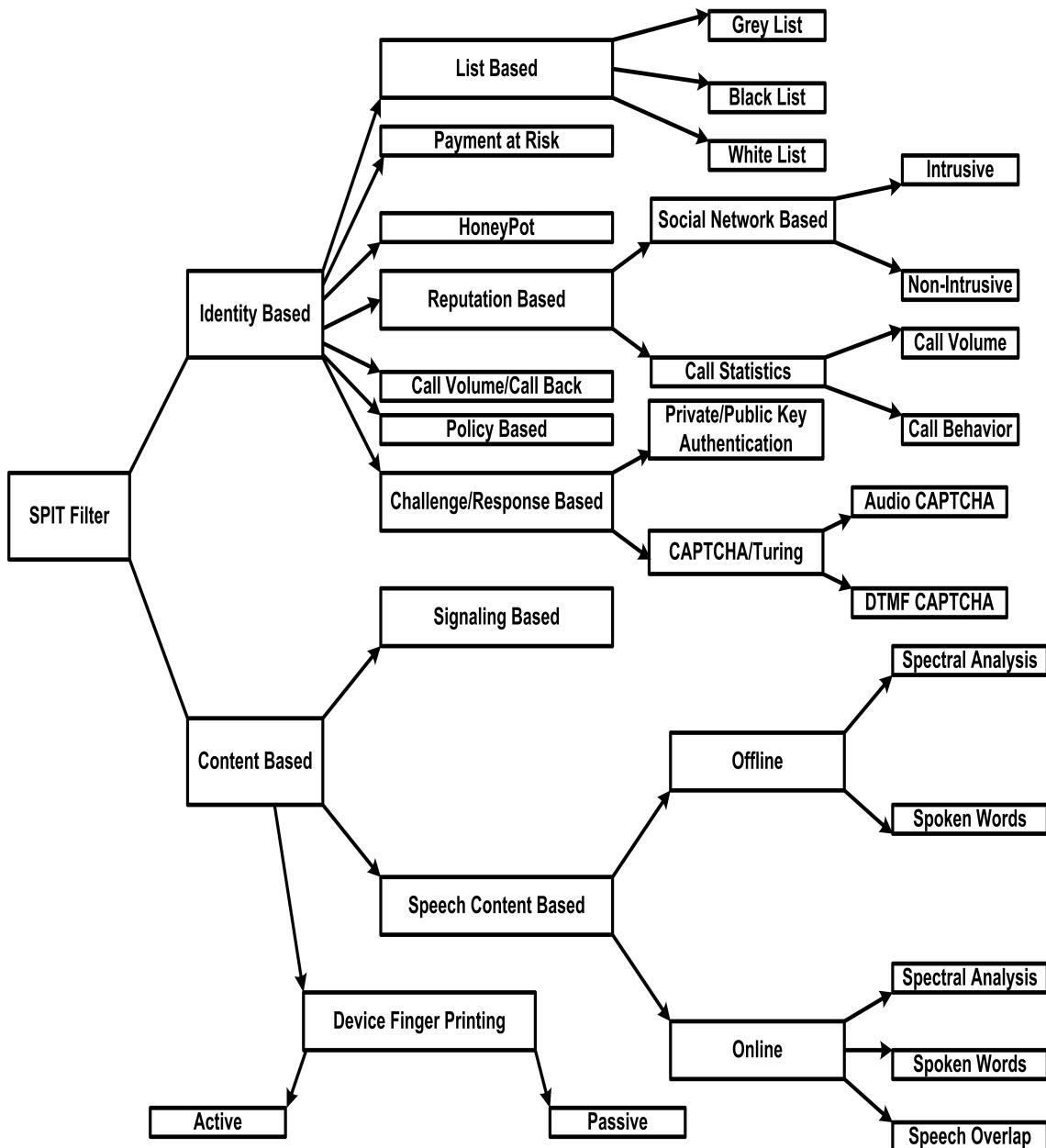


Figure 2.1: A Taxonomy of SPIT Detection Systems.

media server for unattended calls by the caller. The similarity between speech samples can also be estimated by estimating the distance between online speech samples and fingerprints [SMG+12], [LGK+11], [PK08] of speech samples left on media servers. The content-based approaches can also be implemented in the form of multistage systems collaborating with other approaches in order to improve the detection rate and minimize false detection. In [SNT+06], a two stage Multi-layered Fusion-based method has been presented that considered information from signaling messages and speech contents while

making decision about the behavior of a subscriber<sup>1</sup> as a spammer or a non-spammer.

Content-based approaches have shown great resistance against spammers in an email or social networks. In these networks, contents are generally available in the form of text and images that does not require sophisticated system resources for processing and matching. Moreover, users of these networks are normally not worried about the delay incorporated between their conversations when spam detection systems have been deployed for analyzing the contents. However, in a VoIP network, applying content-based systems for spam detection have several limitations because of processing of voice streams in real-time. Firstly, the deployment of speech processing engine would introduce some noticeable delays between conversations of subscribers and this delay greatly annoy the subscribers as they do not want to have long delays in their conversation. Secondly, the service provider requires sophisticated signal processing system and network resources for the processing of online speech streams in near to real-time. Thirdly, content-based approaches decide about caller at a second stage and spammers have already annoyed call recipients with the spam signaling (telephony ringing) and listening to some seconds of call. In the perspective of spammers, content-based approaches can easily be circumvented by slight modification of speech contents and addition of noisy speech streams.

The call setup messages are exchanged between subscribers which are usually available in the form of plain text. The call setup messages can also be processed for the identification of spam content and spammers. However, the structure and messages exchanged during the call setup phase and termination phase does not have valued information to be used for spam detection. Additionally, the call setup messages of spam and non-spam calls are same and do not provide any information to be used for classifying subscriber as a spammer or a non-spammer. It is very easy for spammers to craft the call setup messages so that it looks like a call that is originated from a legitimate caller. In some scenarios, content-based approaches would not have a global implication as people from different regions have different calling and greeting behavior. Lastly and most importantly, in most countries listening, recording and processing speech streams of subscribers is prohibited by law thus the privacy of the subscribers is not protected at all.

### 2.1.2 Challenge/Response-based Detection Systems

A C/R (Challenge/Response) system is a type of spam filtering system that initiates an automatic challenge to the subscriber and subscriber needs to reply correctly to the given challenge. The C/R-based systems allow subscriber to call if the subscriber correctly responds the challenge and block subscriber if he fails to solve the challenge. Spam

---

<sup>1</sup>The terms such as subscribers, end-users and callers are interchangeable in this thesis

call can be generated by the human or an automated machine. Humans (legitimate and non-legitimate) can easily solve the challenge initiated by the proxy server, whereas machines would not be able to solve initiated challenges in a timely manner. The C/R-based anti-SPIT approaches can be implemented in two ways: 1) through authentication and authorization carried out in a non-intrusive way without involving subscriber for the explicit response and instead using encryption or handshake mechanism with the subscriber at the time of registration and call request [SS04] in a seamless way, and 2) a method where the proxy server initiates a CAPTCHA challenge to be solved by the subscribers [LK07], [QNT+08], [QNT+07] in a timely manner.

In [SS04], subscribers are asked to pass through two authentication phases before placing a call to the callee. In the first stage, authentication of subscriber is carried out through digest access authentication – exchange of credentials on which the VoIP proxy server and the subscriber agreed upon such as username or password, and in the second stage, a proxy server authenticates the subscriber through transport layer security and DNS service records. Legitimate subscribers normally have pauses in their conversations, whereas spammers do not have pauses especially at the start of the conversation that results in a speech overlap between spammers and their callees. A hidden Turing test mechanism has been proposed in [QNT+07] that considers talking behavior of subscribers and monitors the overlaps in the speech streams flowing between subscriber and his call recipients. In [LK07], service provider asks subscriber to solve a Human Interactive Proof (HIPs) test by pressing phone keys against initiated challenge. Text and image-based CAPTCHA are not feasible to deploy in a VoIP or voice network, therefore speech-based CAPTCHA is the only available option for initiating the challenge to the subscriber. The subscriber will then respond by pressing combination of certain phone keys in the form of DTMF (Dual Tone - Multi Frequency) signals. In [SG10], audio CAPTCHA is generated to the subscriber and subscriber responds the challenge by pressing phone keys. In [SKE+14], authors combine audio CAPTCHA and game theoretic model for authenticating the subscribers. Besides, solving the challenge and providing the credentials, subscribers are also asked to call back the service provider [CO05] for proving their authentication. The approach presented in [CO05] uses anonymous verifying authority and mediator for initiating call back request and block subscriber if call back request is not fulfilled within a specific time period. A C/R-based system can also be combined with other anti-SPIT systems so that challenge is initiated only to few subscribers or to only new subscribers. The multi-stage C/R-based system would minimize the false detection rate and improve the detection accuracy.

C/R-based approaches are well suited for blocking machine or auto-dialer spammers but deployment in a real-network has problems. Authenticating subscriber via

CAPTCHA or public-private key would introduce notable call setup delay which might displease subscriber for each call made. Though public-private key-based C/R systems are non-intrusive but they require public-private key infrastructure for assigning keys. CAPTCHA test is intrusive to the subscribers and additionally has some other limitation. A CAPTCHA system usually requires solving some challenge which might be difficult to solve by the people with a certain degree of disability. In perspective of spammers, CAPTCHA-based approaches can be circumvented by hiring cheap human spammers. For example, spammers can set-up a network of cheap workers for solving the audio CAPTCHA challenge and then relay voice streams from the media server upon successful authorization of the call. In perspective of service provider, C/R-based system especially CAPTCHA system requires additional network and system resources for initiating and processing of challenges and their responses for a large number of subscribers in parallel.

### 2.1.3 Access List-based Detection Systems

The list-based approaches are the simplest identity based anti-SPIT filters. The proxy server checks the database list during the call setup phase. Whitelist maintains the database of identities that are allowed for using the network and blacklist maintains the database of identities barred from calling. A list database can be either global – applied to all subscribers or the local – applied to the particular subscribers. In addition to black and white list, a grey list [HHM+06], [SAS06] can also be used for maintaining the list of subscriber to be observed for further time period. A greylist maintains the database of identities that are challenged for the authorization or monitored for some extended time periods. In a greylist, if a subscriber exhibits spamming behavior over some pre-defined extended time period then subscriber is permanently moved to a blacklist, otherwise subscriber is moved to the whitelist. The list-based approaches need to be implemented along with other approaches that actually decides which list the user should be placed in [KD07], [DK05], [OS09].

A common problem with blacklist-based SPIT filtering is the inclusion of legitimate subscribers who were mistakenly reported or classified as the spammers because of personal dislikes. Additionally, it becomes difficult for the subscribers to get their identities out of the blacklist. Spammers can easily evade the blacklist database by spoofing identities of legitimate subscribers not included in the blacklist. A whitelist is useful for controlling which subscribers are allowed for calling globally and local. However, it would not allow new subscriber or a legitimate subscriber wishing to call someone for the first time. In perspective of service provider, maintaining a large global and local list database is problematic and requires continuous update. Moreover, list-based approaches need to

be implemented with other SPIT detection approaches that actually decides whether to include subscriber in the respective (black,white and grey) list or not.

### **2.1.4 Cost-based Detection Systems**

VoIP offers affordable calling rates that has not only attracted legitimate subscribers but has also attracted spammer for making unsolicited advertisement calls to a large number of recipients. One way to block spammers is to impose some extra cost on subscribers classified as spammers. In this scenario, a service provider first deducts some money from the subscriber credit and returns back if subscriber is classified as legitimate later at end of call. If subscriber is classified as a spammer later on then the service provider will withhold the deducted money and distributes this money among spammer's victim callees. The cost-based approach needs to be implemented with other approaches [RJ08] that actually decides whether to return back money or not.

A major limitation of cost based system is that it requires comprehensive micro-payment system for the computation of deduction and holding of money. Moreover, cost-based systems are dependent on other SPIT classification methods for the final decision. It might also be possible that cost-based systems would charge legitimate subscriber and would also mistakenly block legitimate subscriber if subscriber has insufficient balance required for the deduction but greater than the amount required for the call.

### **2.1.5 Policy-based Detection Systems**

In a policy-based anti-SPIT systems, service provider defines policies that instructs anti-SPIT system to monitor the behavior of incoming call request according to subscriber's network-wide policies [GKB+12]. In [DSN09], authors proposed policy-based anti-SPIT system that maintains subscriber preferences and policies using call processing policy framework [STM] for blocking the spammer. In [SDG08], authors used adaptive policy system based on the definition of a set of rules along with actions and controls for mitigating the SPIT attack. In [TFP+06], authors presented a system that uses Security Assertion Markup Language (SAML) for authenticating the subscriber in the network. The policy-based systems are well suited for the small operators or local enterprise organization because maintaining policies for thousands of users in a big operator are difficult and unmanageable task.

The major limitation of policy-based SPIT detection systems is the maintenance and updating of subscriber public and private policies. Moreover, policy-based systems would

add noticeable delay during call setup phase because of processing of policies of subscriber and his behavior towards these policies. In perspective of spammer's attack, policy-based system can be circumvented by spammer through spoofing the policies of the legitimate callers.

### 2.1.6 Legislation-based Detection Systems

The primary goal of the legislation is to create a legislative framework that would make spamming illegal and impose punishment on those involved in spamming activities. European Union, USA and Canada has already made some reasonable efforts in terms of legislation against initiators of spamming [ITU15], [CJE15], [LAW15]. These legislations prohibit unsolicited communication to reach the recipient unless prior consent of the recipient is obtained. The major limitation of legislation-based anti-SPIT system is the difficulty of tracing back the initiators of spam communication for the law enforcement agencies. Moreover, if the regulator or law enforcement agencies trace-back the initiator of unsolicited communication even then there is no such global law exists that will apply to spammers across the world. Moreover, spammers make spams from anywhere around the world thus make anti-spam law of one country inapplicable to the spammer spamming from places where no such law exists.

### 2.1.7 Multi-Stage Detection Systems

The Multi-Stage anti-SPIT systems require internal collaboration among many independent standalone systems within the service provider in order to improve the detection accuracy and detection time. The multistage systems collectively utilize information from many standalone components or systems in order to make decisions about behavior of the subscriber. In [SNT+06], authors presented two stage system that processes speech streams and signaling messages of the subscriber for blocking the SPIT caller. Similarly, in [GM08], authors presented a two stage system that is based on the CAPTCHA test and a speech processing. In [DK05], authors presented content independent multistage system consisting of three stages: 1) a Rate limiting stage – that monitors call statistics such as call rate in a certain time period, 2) a blacklist module – for including spammers in a blacklist, and 3) a multivariable Bayesian network – for inferring reputation of a subscriber from the subscriber direct trust scores with other subscribers. In [QNT+08], authors use different thresholds for computing reputation at each stage of multistage detection system. In a first stage, the system computes reputation score of the subscriber and

compares this reputation score against two thresholds – low and high thresholds for forwarding call to the next stage. In a second stage, the system invoke Turing test for authorization and in a third stage the system asks call recipients for the feedback about the initiator of the call. In [AM11], authors presented a four stage system for blocking spammers in a transit VoIP network using internal collaboration between many stages. In [SAS06], authors use subscriber's short and long time call patterns and blocks spammers using collaboration among various list databases – white, black and grey list. In [MNS08], authors proposed multistage system that incorporates feedback from multiple stages while making decision about subscribers as a spammer and a non-spammers. The multistage stage system requires collaboration among well-known detection systems such as blacklists, whitelists and call statistical analysis (call duration and call rate) into a multistage SPIT detection system.

It is obvious that multistage systems would provide better detection accuracy and detection time but their performance depends on the types of standalone systems used together. The content-based multi-stage systems have same limitations as of approaches based on the speech content processing. Similarly, the CAPTCHA and reputation-based systems require interaction with the caller and the callee, thus are intrusive. Another major issue with multi-stage systems is that they would introduce considerable high delay during call setup phase.

### **2.1.8 Call Statistics-based Detection Systems**

A VoIP call consists of two parts: 1) a signaling phase – a series of call setup message exchanged between subscribers and the proxy server at the time of call request, and 2) a speech streaming phase – the exchange of actual speech content between caller and the callee after the call establishment phase. The statistics-based SPIT detection systems monitor different call statistics of the subscriber that are: the call-rate, the call duration, inter-arrival time between call requests made by a subscribers, number of speech packets exchanged between subscribers, which subscribers disconnected the call etc. Several machine learning approaches have also been applied to call statistics and calling behavior of subscribers for differentiating spammers from the non-spammers [TS13]. The information from call statistics (call rate, call duration etc.) and calling behavior (number of friends, number of incoming calls etc.) of subscribers can also be used for computing reputation of the subscribers which is then used to block subscriber if reputation of subscriber is less than certain threshold.

The major limitation of statistics-based approaches is learning the behavior of legitimate and non-legitimate subscribers. It would be difficult in a statistical system to differ-

entiate the spammer from the non-spammer with small false positives. In perspective of spammers, statistics-based systems could be easily circumvented by spammers by controlling the similar call statistics and also by spoofing identity of legitimate subscribers. In perspective of deployment, learning a dynamic threshold for each subscriber and for each aggregation cycle is also challenging and problematic.

### 2.1.9 Device Fingerprinting-based Detection Systems

A device fingerprint is the information that has been recorded on a proxy server or a remote computing device for the purpose of identifying devices and software used by the subscriber. The fingerprinting-based anti-SPIT systems create fingerprints for a set of devices used for making calls in a VoIP network. The fingerprinting-based approaches assume that spammers and non-spammers uses different telephony devices and communication protocol stakes for making and receiving calls. These approaches are grouped in two types: active fingerprinting and the passive fingerprinting. In active fingerprinting remote device asks subscriber for transmitting packets to remote system for device analysis, whereas in the passive fingerprinting remote device actively monitor the fingerprints of device originating the calls. In [HSZ+06], authors analyze fingerprints of several commercial hard and soft phones for different call response messages using active and passive fingerprinting. The use of device fingerprinting in real deployment is not practical and scalable as it requires management of fingerprints large number of commercial and non-commercial VoIP devices. Additionally, spammer can bypass fingerprinting-based systems by adopting fingerprints and protocol stack similar to the devices used by the legitimate subscribers.

### 2.1.10 Honeypot-based Detection Systems

Spammers normally crawls web or telephone directory for the collection of target identities without knowing whether target identities are real or virtual. Honey phones are virtual phones not assigned to human users but are used for analyzing the behavior of subscribers making calls to them. As honey-phone is not assigned to any physical person thus naturally would not make any call to users or receive any call from the legitimate subscriber. Subscribers calling honey-phones could be spammers but confirmation requires analysis of subscriber behavior towards honey-phones [NNS+07], [GSB+15], [LCW10] and other network users. In [GSB+15], authors deploy a large scale cloud based honeypot system for analyzing the social behavior of callers making calls to these honey-phones. In [BGG+16] authors deploy a MobiPot - a honeypot mobile system for collecting the

fraudulent calls and SMS. These calls and SMS are then analyzed for studying the mechanism used by spammers for collecting the identities of target victims and spamming attack patterns.

Honeypot-based solution can identify spammers that are targeting to them but would not be able to identify those spammers spamming other users in the network. The spammers can also bypass honeypot systems by learning the numbering pattern of honey phones or by using phone numbers of confirmed human beings.

### **2.1.11 Reputation-based Detection Systems**

Collaboration among subscribers can assist subscribers who wish to make decision about whether to receive or reject the call from the subscribers not known to them. The reputation-based systems operate in two ways: distributed – where each subscriber directly collaborates with other subscribers and a centralized system – where subscriber directly collaborates with the centralized service provider or system. Several reputation-based anti-SPIT systems have been proposed for filtering spam in a VoIP network. These systems consist of two steps: computing direct trust between subscribers engaged in communications and then aggregation of direct trust scores of subscribers for their global behavior. The direct trust represents strength of direct relationship between subscriber and his interacted subscriber, whereas the global reputation represents aggregate behavior of subscriber towards all his interacted subscribers. If trust and reputation score of the subscriber is higher than some learned or fixed threshold then subscriber is considered reputed otherwise subscriber is considered non-reputed.

The direct trust between subscriber and his interacted subscriber can be computed in two ways: 1) intrusive way – that implicitly requires interaction with the call recipient of the subscriber [KD07], [DK05], [WBS+09] for the feedback about subscriber, and 2) a non-intrusive way – that explicitly utilize information from call logs recorded for the billing purposes [BAP07], [RSM06], [RS05]. The global reputation score of the subscriber can be computed in two ways: by applying Eigen Trust [BAP07] to the direct trust scores and by applying machine learning approaches such as Bayesian networks and clustering to the direct trust scores of subscriber [KD07], [DK05],

Spammers normally exhibit different calling behaviors from the legitimate subscriber with the following properties: they target large number of recipients, receive calls from only few callees and many of their received or made calls are of small duration. This calling behavior normally results in a spammer's disconnected social network with the large number of their target victims. On the other hand legitimate subscribers normally have small number of recipients, have repetitive calling behavior with many of their called

callees and also receive good number of calls from their called callees. This calling behavior of legitimate subscribers results in a subscriber's strong relationship network with many of his callees. In [BAP07], authors proposed a Call-Rank system – a reputation-based system that uses average call duration for computing direct trust between subscriber and his called callees. The global reputation of subscriber is then computed by applying Eigen trust algorithm to the subscriber's direct trust scores. Finally, Call-Rank asks callee for the final decision (accepting call or rejecting call) by sending reputation scores and social network credentials of caller to the callee. In [ZG09], authors considered subscriber interaction with his callee as reputed if call duration is greater than 20 seconds. In [COB+11] authors used call duration along with seven degree separation as a social network feature for the computing of reputation of the subscriber across the network. In [BSG+11], authors proposed three reputation-based solutions for filtering SPIT subscriber. In the first approach, they used concept of strong and weak social ties among subscribers, in the second approach, they enhanced Progressive Multi Grey-Leveling list to the Enhanced Progressive Multi Grey-Leveling by using call density and reciprocity-index features, and in the third approach, they adopted Page-Rank algorithm for computing global reputation of the subscriber.

The direct trust between subscriber and his callees can also be computed by collecting feedback (positive or negative) from the callee for the subscriber's call transaction which just ended. In [WBS+09], authors computed reputation of the subscriber aggregating callee's feedback about the subscriber's calls. The reputation scores and call statistics of subscriber are then used along with MPCK-Mean – a semi supervised clustering algorithm for clustering subscriber into SPIT and non-SPIT clusters. The proposed system performs well only when callee provides honest and accurate information about the caller. In [KD07] and [DK05] authors proposed a multistage system for the identification of spammers. The system consists of three stages that collaborates with each other: the direct trust stage computes direct trust between subscriber and his callees by aggregating feedback from the callee for the subscriber call, the global reputation stage that computes reputation by applying Bayesian network algorithm to the direct scores of the subscriber to all his callees, and the list database that maintains list of black and white listed subscribers. In [PD09], authors proposed an approach that aggregates callee feedback along with G-mail spam filtering method for blocking the spammers. In [PGK+08], authors proposed two approaches based on the content processing and feedback aggregation about behavior of the subscriber from his callees. In [SDN+09], authors computed reputation of the subscriber through a web of trust model between subscriber in a network. In [GYH08], authors proposed a multilayer system that incorporates behavioral characteristics of subscriber and his call signaling messages. One of the major limitations

of intrusive reputation approaches is their intrusiveness and also requires change in a VoIP handset and call signaling messages. Moreover, spammers can easily circumvent intrusive approaches by creating network among his identities and providing fake responses for their identities.

Reputation based anti-SPIT systems have shown great effectiveness against spammers in email and VoIP network but their effectiveness depend on the set of features used for the computation of global reputation. In some cases, spammer could get high reputation scores by creating a Sybil network between his acquired identities and also spoofed identities of the legitimate subscriber for spamming and getting the high reputation scores. The CDR based-reputation systems minimize the effect of Sybil attack but its performance depends on features used for the computation of global reputation scores. The spammer normally targets large number of callees without repeating his callees thus normally results in a small duration calls to a large number of callees and good duration calls to only few callees. In non-repetitive calls, the average call duration of caller with the callee is same as of his aggregate call duration. This behavior might results in a high reputation scores for the subscribers having good duration calls to a large number of their called callees. Moreover, spammers can also collude among their several identities with good duration calls that would also increase spammer's global reputation. In some CDR-based approaches, the global reputation score and social network credential of the caller are also sent to a callee for the final decision [BAP07] which is not only intrusive to the callee but also poses threat to the privacy of the caller.

## **2.2 Collaborative Detection Systems**

Spammers always try to find ways for evading the spam detection systems. They manipulate contents by adding noisy messages, acquire large number of identities for colluding and Sybil attack, and controlling number of spam calls to a single service provider but target many service providers in parallel. Spammer can make a large number of spam calls in aggregate from a given calling identity but distribute calls among many service providers. Existing anti-SPIT system decides about behavior of the subscriber based on his calling behavior observed at a single service provider. These approaches prolong detection time and block spammers only if spammers make significantly large number of spam calls to a single service provider. The non-availability of calling behavior of spammers across the service providers limits standalone system to react effectively against spammers.

The calling behavior of subscriber becomes more meaningful when subscribers are observed across many service providers. Naturally, collaboration among service providers

would improve the detection time and detection accuracy because of collective use of information about behavior of subscriber from many autonomous collaborating service providers. Service provider or domain collaboration has been applied in various networking domains for identifying malicious intruders and spammers in a network. In this section, we provide detailed discussion on the collaborative systems in the perspective of SPIT and email-spam detection systems.

### 2.2.1 Collaborative SPIT Detection

In order to evade the standalone systems, spammer targets large number of recipients that are dispersed across many service providers but his calling behavior remains same across all service providers. Collaboration among service provider is a natural way for early detection of spammer before they spam a large number of recipients. However, a very few works have been reported that incorporate collaboration among service providers for rating the subscribers. In [IMS2015] and [3GPP2015], 3GPP a standardization body on next generation network formalized best practice standards for fighting spammers in a VoIP and IMS networks. Particularly, they encouraged service providers to have collaboration among themselves for early and effective IM (Instant Messaging) and voice spam detection. However, this technical standard does not provide any information about how collaboration is to be carried out. In [SLS+10], authors exchange scorecards of subscriber among collaborating service providers and collaborating service providers react independently against the subscriber's scorecard whether to allow or block the subscriber. This framework requires predefined trust relationship between collaborators which is practically not feasible in a telecommunication network. Additionally, no mechanism has been presented for the computation of subscriber score and trust assessment of collaborators.

In [SS09], authors proposed a collaborative system where the home service provider collaborates with the visiting service provider with the exchange of information about their SPIT detection system in the form of call tags. The visiting service provider then access the performance of SPIT detection system deployed in a home service provider of the subscriber calling recipients of his network. This approach has some limitations: 1) it only evaluates performance of a SPIT detection system deployed in a service provider that provides the tag information, 2) it requires establishment of predefined trust between collaborating service providers, and 3) it requires change in the call signaling messages exchanged between collaborating service providers in order to incorporate tag information. In [WAB+09], authors proposed SPACEDIVE that detects intrusion in a VoIP network by correlating local and remote information from many collaborating service providers. In [WMH07], authors presented P2PAVS that computes reputation of subscribers through

collaborative response from the subscriber from many service providers. However, it does not provide any mechanism for the propagation of trust among subscribers. In VoIP normally collaboration is achieved in the form of multistage systems or collaboration among proxy servers within the service providers. In [SKE+14], authors proposed an approach that uses collaboration among several local VoIP servers within the service provider. However, mechanism for collection and aggregation of feedback among subscriber and proxy server is not provided.

### **2.2.2 Collaborative Spam Detection in Email Network**

A number of collaborative systems have been proposed for filtering spammers in an email network and detection of intruders in the IP-based networks. In perspective of email networks, the collaborative systems normally require collaboration in the form of exchanging spam message contents, spam HTML (Hyper Text Markup Language) tags and exchanging feedback about behavior of particular sender. Normally spammers send same spam message to a large number of recipients and collaboration with the message content would greatly improve the spam detection time but it poses threat to the privacy of email users. In [LZR09], authors presented a privacy-aware collaborative system called ALPACAS that invokes collaboration among service provider with the exchange of encrypted fingerprints of message contents. The similarity between fingerprints of user messages and messages stored in the database of ham and spam is estimated in order to classify the new message as a spam or a ham. HTML tags are available within the email headers and can be used to distinguish spam content from the non-spam. In [YPC+11], authors presented a COSDES system that collaborates with the exchange of HTML tags and computes distance between spam and ham tages using near duplicate approach. In [FZN06], authors presented a collaborative framework that requires distributed collaboration for making decisions about users. The collaborating email domain directly exchange information to each other and imposes some restrictions on domains not taking part in a collaboration process.

The behavior of spammer remains same in all target service provider or domains and when analyzed collectively would decrease the detection time. In [RFV07], authors proposed a system called SpamTrackers that requires collaboration among domains with the communication behavioral patterns of email senders within collaborating domains. In a [DVP+04], authors presented a three layerd P2P architecture based on communication patterns of spammers and non-spammers. The architecture requires collaboration among end-users, mail service and the super peers. The super peer handles the exchange of message among themselves for tagging and classifying incoming mail digest as a spam or a

non-spam. In [SBK07], authors presented RepuScore that require collaboration among email domains with the exchange of local reputation score of email sender within the domain. In RepuScore, a centralized system computes global reputation of user by aggregating local reputation scores. In [SKY11] and [CLO16] authors proposed social filter, a centralized system that aggregates feedback from individual spam detection systems for early detection of Phishers and spammers.

Collaboration can also be carried out directly among end-users. In [KRS+06], authors presented a collaborative spam filter that uses collaboration and social network information of users for blocking spammers. In [CDN05], authors proposed two spam detection systems: a simple Mail-Rank and a personalized Mail-Rank that computes global reputation of email user by aggregating direct trust score through power iteration algorithm. Spammers are moving to different online social networks for increasing their footprint. The collaboration among different social network platforms would greatly improve the detection accuracy and detection time. In [WIP11], authors proposed a system that incorporates collaboration among different online social network platforms with the exchange of spam contents to be used for effective spam detection.

All of the above proposed collaborative approaches classify incoming email into spam or non-spam by analyzing the contents of messages, contents of HTML tags and static rule for some features. These approaches cannot be applied directly for filtering spammers in a voice networks. In a voice network, contents are available in the form of speech signals and having collaboration with the exchange of speech signals is not feasible. Moreover, it requires sophisticated system and network resources for speech processing, storage and matching. In case of non-content-based collaborative approaches, the detection approaches utilize structure of the network while ignoring the weights on the links between users. In a voice network, few additional features such as call rate and call duration could provide information about relationship strength among users and should be used in a collaborative way.

## 2.3 Identity Linking

There are many countries where number of subscribers exceeds the country population. For example Russia has 1.8 percent more mobile subscriber than its population, similarly Brazil has 1.2 percent more subscriber than its population. These numbers are not attributed to the fact that every citizen has one identity but attributed to the fact that many individuals have more than one calling identity. As VoIP offers cheap telephony rates

and acquiring new identity in a mobile and VoIP network is not costly that's why spammers are exploiting VoIP network with the spamming activities from many identities. The spammer normally makes controlled attack from all his identity or rejoin network with new identity if blocked by the detection system. Moreover, spammer's identities also collude with each in order to have high reputation score so as to remain undetected. The standalone and collaborative SPIT detection systems are able to identify such spammers that are making large number of calls from their each identity. The spammer can change his identity once its blocked by the service provider and targets user with new identity with same motive. However, connecting or linking multiple identities that belong to one physical person would greatly decrease the detection time and would improve the detection accuracy. Identity linking would also help in characterizing the complete behavior of physical user having multiple calling identities for other purposes such as recommendation and marketing of other value added service. To the best of our knowledge, we have not found any such study that has been carried out for linking multiple calling identities of the a physical individual in a voice network. For this reason, we are providing works that have been done for linking multiple identities of user across same or different online social networks.

Normally, an active social network user or Internet user has multiple accounts across different social network or same social network with same or different identity. Current statistics show that 20% of facebook users also have twitter account and 91% of Twitter users also use Facebook to stay in touch with others [PER09]. In online social network like Facebook, twitter or Instagram etc. profiles are linked together by estimating similarity between identities in three dimensions. Firstly, profiles can be linked together by matching the profile information provided by the user at the time of creation of account across different social networks; secondly, profiles are matched by estimating the similarity in content posted by the user across different social networks; and thirdly, profiles are linked together by using information from friend's network of users from different social networks. In [RCD10], authors considered three dimensions and presented an identity linking framework that links similar profiles together that belongs to one physical individual. In [VHS09], authors proposed a system that links profiles of users by estimating the similarity in profile information represented in a profile vector. Moreover, authors also identify number of significant features that better characterize the profiles of the same user. In [NCM12], authors represented user's profile information as a feature vector and applied supervised machine learning (Support Vector Machines, Random Forests and Alternating Decision Trees) for making decisions about identities that belongs to one physical individual.

Users usually exhibit unique communication behavior which they repeat across different social networking sites. In [ZL13], authors considered behavioral patterns of users and proposed a MOBIUS system for finding a mapping function among profiles of similar user from different social networking sites. Primarily, MOBIUS consist of two important components: first component uses behavioral patterns and the second component employs machine learning approach for user identification and linking of the same profiles. In [ZL09], authors analyzed naming patterns of users across 12 different social networks for finding features that better link accounts that belong to the same individual. In [SCA08], authors consider tagging behavior for connecting user's flicker profile to his del.icio.us profile. In [IFA+11], authors linked similar profile by processing information in two aspects: user ids and user tags.

The collective use of social network features, profile features and content features would greatly improve the linking accuracy. In [JKJ13], authors collectively use the content and network connection of user to connect his Facebook profile with his MySpace profile. In [ACF13], authors used partial social network of user semantic similarity between his profiles attributes for linking his identities. In [MLM+12], authors applied jaro distance and TFDF vector space model to user's profile information and his social network. In [GLP+13], user's accounts across social networks are correlated using innocuous information such as location, content posted across social networks, writing style of posted content and time when content posted. In [LWZ+14], HYDRA system applies multi-objective optimization (MOO) to the heterogeneous behavior of users and their core social network structure. The links between users can also be predicted by applying supervised unsupervised link prediction algorithms to the user's social network attributes. [GTM+14]. In [BKP12], JAL system uses Conditional Random Fields to user's profile and his social network information for resolving identities of users across different social networks. In [ZLC+06], a framework is presented that considered four type features from messages for the authorship identification. In [SSB05], several similarity measure approaches are compared and evaluated for recommending online communities.

Several works have also been proposed for estimating the node similarity in social graphs through use of node's social connections. The node connections are basic building blocks for link prediction, collaborative filtering, identity linking and record linking. The similarity among nodes of different graph or same graph can be computed local or global: local measure estimates similarity between nodes using information from immediate neighborhoods (for example number of common friends between nodes) of nodes, and global similarity computes node similarity considering complete graph structure of nodes. The similarity between graphs and graph nodes can be computed by iteratively adding the similarity scores between neighbors of nodes in a both graphs [BGH+04],

[MMR02], [FZX15]. It is important to consider over all structure and connection of nodes that is friends of friends of node and has been adopted in SimRank [JW02] system used for link prediction and node similarity. In [HGL+11], ReFeX (Recursive Feature eXtraction) system recursively aggregates regional features from the neighboring nodes. These regional features represent the type of nodes to which a given node is connected to.

The identity linking problem can also be seen in terms of de-anonymization problem. In [NS09], authors proposed a de-anonymization approach that identifies nodes using structural information of nodes in a social graph. The algorithm requires few starting seeds and propagates mappings through node edges. In [BKE+13], authors presented a NetSimile approach that incorporates theories from social network for estimating similarity between nodes of two graphs.

Many researchers have applied identity linking to the problem of spam detection on social networks and email networks. In [XYY+13], authors proposed a VoteTrust system that combine network structures of individuals and their feedback to detect individuals having multiple identities. In [XFH15], authors presented a machine learning approach that uses interaction patterns and profile information of users for detecting fake accounts in online social network. In [JTJ11], authors proposed an approach that analyze the behavior of identity based on attribute similarity and similarity in friendship network to detect the fake accounts.

## 2.4 Discussion

To date, existing anti-SPIT systems have taken two main mechanisms: 1) content-based filtering and 2) identity-based filtering. If not properly designed, both mechanisms have some limitations and can be evaded by the spammers with new ways. In Section 2.1, we outlined existing standalone anti-SPIT systems and their limitations. Content-based system could block certain type of spammers but they can be circumvented by spammers through slight modification of speech content. Moreover, applying content-based approaches in a VoIP for filtering spammer has some additional limitations as discussed in section . Identity-based reputation systems are viable for blocking spammers in a VoIP network but these systems can be evaded by the spammers if approaches are not carefully designed as discussed in section 2.1. The challenge is to design a content independent, reputation-based and non-intrusive SPIT detection system that does not require any change in existing VoIP network and handset of subscribers.

A number of reputation based systems have been proposed [KD07], [DK05], [WBS+09],

[BAP07] but all are intrusive to end-users. Call-Rank [BAP07] though computes subscriber's reputation automatically but it relies on the callee for the final decision and has other following limitations as well. Firstly, it discloses subscriber's reputation scores and social network credentials to the callee which could be threat to the privacy of the subscribers. Secondly, it requires public and private key infrastructure for authentication and authorization. Thirdly, it asks subscriber to solve the CAPTCHA challenge to minimize the false positives and handle calls from new subscribers. Fourthly, spammer can easily evade this system by creating multiple identities and developing strong relationships. However, we believe that anti-SPIT systems must be non-intrusive and should collectively use number of call and social network features. Moreover, classification threshold needs to be computed automatically without incurring high false positives and small true positives. The collective use of social network and call features in both directions – incoming and out-going could greatly improve the detection accuracy and reduce interaction with subscribers. It also becomes difficult for spammers to evade such system because they are not able to control multiple social and call features.

The standalone system either based on the reputation or the content prolongs detection of spammers if spammers make small rate spamming to the recipients of one service providers but target recipients of many service providers. Collaboration among service provider is the natural phenomena for blocking these small rate spammers in a timely way. However, to-date, no collaborative system exists where telecommunication service providers collaborate with each other by exchanging information about behavior of subscribers in their network. Moreover, service providers are not willing to be part of collaboration with each other because they are business competitor and worried about privacy of their customers. The challenge in the design of a collaborative system is to convince telecommunication service provider for taking part in the collaboration process. The simplest collaborative approach to convince service provider is to use the trusted centralized system so that service provider directly collaborate with the trusted centralized repository. However, service provider requires absolute protection of privacy of their customers even in presence of trusted centralized repository, they are only interested in exchanging such information which does not contain any threat to privacy of their customers and its network configurations. There is strong need to have collaborative system that has following features: 1) it incorporates use of trusted centralized repository for aggregation of information exchanged by autonomous collaborators, 2) it uses filtered non-sensitive information from the collaborators, and 3) must protect privacy of the subscribers relationship network i.e. an intruder at a centralized repository would not be able to infer the relationship network of the subscribers.

Spammers can acquire large number of identities because of cheap telephony rates and

easy integration of telephony with Internet technology. These identities either colludes with each other for the high reputation scores so to make spam calls to other users or use new identity once blocked by the service provider. However, spammer having multiple identities has overlap in target identities among his different identities exhibit similar calling behavior towards many target identities. The standalone reputation based anti-SPIT systems react very slowly towards these spammers having large number of identities. It is of utmost importance to link the multiple identities that belong to one physical person so as to collectively use the information for computing reputation of the subscriber. The identity linking is not only beneficial for timely identification of spammers but would also be helpful in characterizing the complete behavior of legitimate subscriber having many identities. To-date, no anti-SPIT system exists that computes reputation of the physical individual by connecting all his multiple identities that belong to him. Moreover, to the best of our knowledge, there exists no such system that links identities that belongs to one physical individual using individual's behavioral and call features in a telecommunication network.

The limitations of existing anti-SPIT systems mandate us to design a behavioral-based collaborative SPIT detection system that identifies spammers without relying on the callee feedback or content analysis. The proposed detection system is able to achieve number of objectives. 1) It uses number of call and social network features while computing reputation of the subscribers and make decisions without involving subscribers at any stage of the call processing. This is merely achieved by investigating and selecting the social and call features extracted from the call logs recorded by the service providers. 2) It incorporates privacy-aware collaboration among autonomous service providers for the early identification of spammers making low rate spamming activities to subscribers of many service providers. This is achieved through the exchange of filtered information that would not allow intruders or adversary at centralized repository to learn the relationship network of the subscribers. 3) The proposed system incorporates process of identity linking to connect the identities that belong to one physical individual or subscriber for identifying spammers who frequently change their identities. This is achieved through the use of call and social network features for linking similar identities of the single individual and then reputation of an individual is computed rather than reputation of identities. 4) A model is presented that generates a synthetic data-set for evaluating the detection system. This is achieved by modeling the social behavior of spammer and non-spammer for different graph networks. 5) Finally, a privacy breach analysis is performed for the anonymized data set and presented a method as an absolute protection of privacy of user when data is outsourced for research and analysis.

# **Chapter 3**

## **SPIT Detection System Based on Social Reputation**

In this chapter, we present an overview of our proposed system for mitigating spammers in a VoIP and voice networks. This chapter first briefly overviews VoIP technology and protocol used in a VoIP telephony. Second, it briefly describes proposed system architecture and provides details about the information extracted from the call detail records used for the construction of weighted call graph among subscribers. Finally, it briefly discusses and analyzes behavior of spammers and non-spammers for the different call and social network features.

### **3.1 VoIP (Voice over IP)**

VoIP enables voice and multimedia communication to be carried over the Internet Protocol rather than on the traditional Public Switched Telephone Network (PSTN). Enterprises and voice service providers are adopting VoIP technology as a preferred medium for the long distance cheap telephony across the globe. A typical VoIP session has two phases: a signaling phase and a voice transmission phase. The signaling phases are responsible for the call set-up and tear down of established calls, and also perform other call management functions. The speech exchange phase is responsible for the transmission of digitized voice between end-users. The two major protocols used for establishing the calls in VoIP network are H.323 and SIP (Session Initiation Protocol). A typical SIP based VoIP network is shown in a Figure 3.1 and consists of variety of network devices that includes, a VoIP proxy server for handling call requests between end-user, a registration server for users registrations, a radius server for the authentication and authorization, and a billing

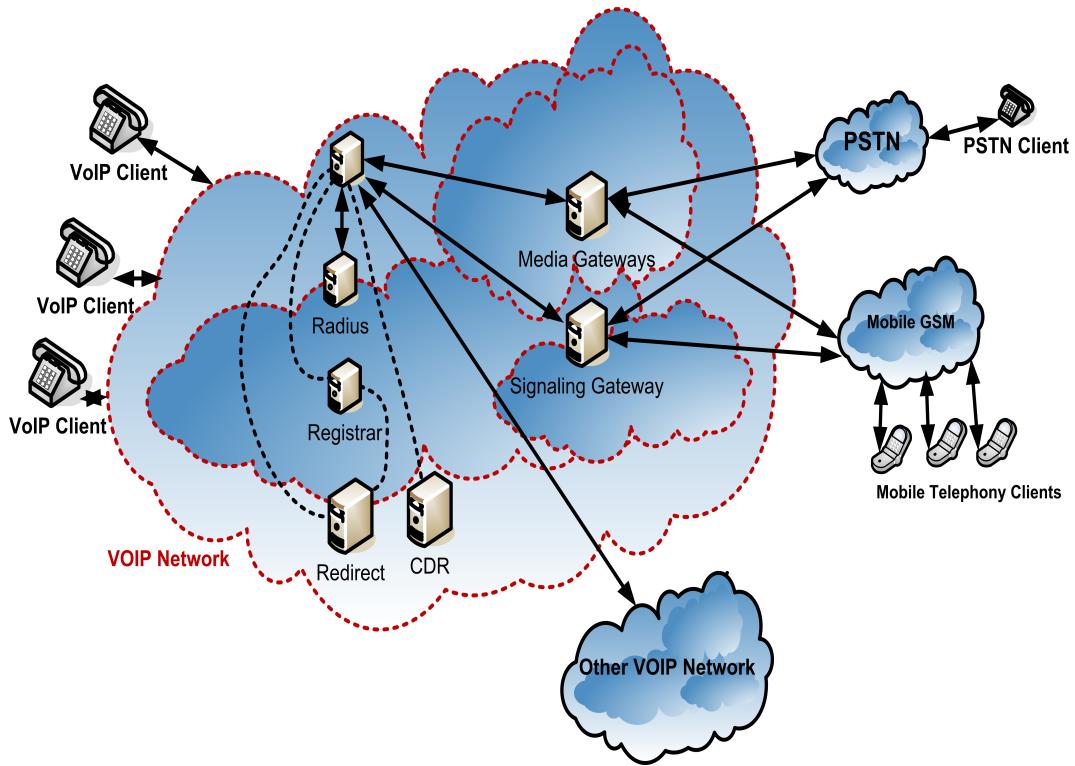


Figure 3.1: SIP Based VoIP Network.

server for logging call detailed records to be used for the billing purposes and network management.

### 3.1.1 H.323

H.323 is a suite of protocol defined by ITU (International Telecommunication Union) as a standard protocol for voice communication over the Internet protocol. H.323 based VoIP network consists of four major components: end user Terminals, Gateways, Gatekeepers and Multipoint Control Units (MCUs). The H.323 protocol suite consists of three major control functions: H.225 for the RAS (Request, Admission and Status) signaling that controls call between end users, a H.245 for the transport of media and signaling message and a H.235 for the secure communication between end users.

### 3.1.2 Session Initiation Protocol (SIP)

The Session Initiation Protocol (SIP) is an IETF standard and the most popular protocol used for managing the calls between users of the VoIP and NGN (Next Generation Networks) network [RFC3261]. The exchange of SIP messages between end-users determines the capabilities and location of communicating entities before exchange of actual

speech signals between them. The SIP based VoIP network consists of two major components: the SIP User Agent (UA) and SIP Network Server. The SIP UA is the user soft or hard phone responsible for initiating and accepting calls. SIP Network Server manages signaling sessions among participating entities and consists of three main functional components: the SIP Registrar, the SIP proxy server and the SIP redirect server. SIP exhibits request and response model for the session management among communicating entities. Request messages are sent from the user to the Registrar server for the registration, call request messages are exchanged between end users for the start of new session, updating the parameters of existing session, acknowledging session establishment between users and terminating the existing sessions [RFC3261]. Response messages are used for providing the appropriate reaction to the request messages, depending on the type of request message. Figure 3.2 represents the exchange of signaling messages used for call management between end-users and core VoIP network. The SIP based VoIP network also consists of other supporting servers such as: a CDR server for storing Call Detailed Record of users call transactions, billing system for billing and presence servers for storing the location and status of users.

In addition SIP and H.323 protocols, VoIP networks may use other protocols establishing, terminating and managing the sessions. These protocols include: SCCP (Skinny Client Protocol) – a CISCO proprietary protocol used by CISCO IP phones and CISCO call manager, a MGCP (Media Gateway Control Protocol) – for controlling media gateways on a VoIP network and an ITU MEGACO (Gateway Control Protocol) for providing interconnection between traditional public switched telephone network (PSTN) and modern packet networks. For the transport of speech streams or voice signals, The Real-time Transport Protocol (RTP) and Real-time Transport Control Protocol (RTCP) are widely used in a VoIP network.

VoIP transports signaling and voice over an IP network thus vulnerable to the security threats already affecting IP network. These attacks include: Voice Phishing (Vishing), VoIP Spam (SPIT), scanning operator's configurations for toll fraud, Dos and DDoS attacks, billing attacks etc; and not only affect the performance of VoIP network but also causes serious discomfort to the end-users.

## 3.2 System Overview

The architecture of our proposed collaborative, social reputation-based anti-SPIT system is shown in the Figure 3.3. The system consists of three parts: the standalone detection system, the collaborative system and the identity linking system. Of three parts, two

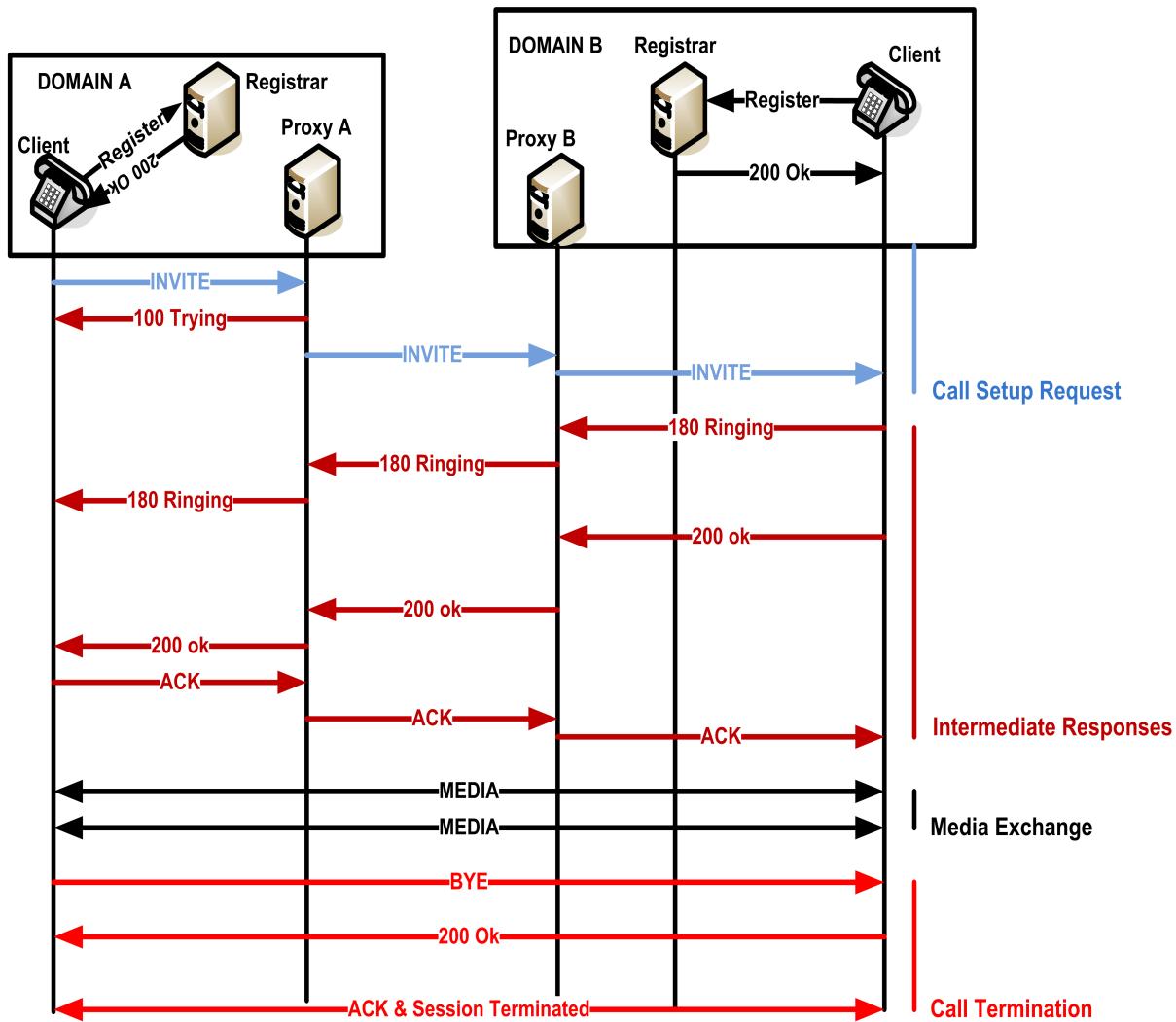


Figure 3.2: SIP Session Establishment and Termination.

parts operates independently, whereas the third part requires collaboration among independent standalone systems placed in different service providers. The intuition for this design choice is based on three observations. 1) Spammers and non-spammers normally exhibit different calling patterns towards their callees. We make use of this difference in the calling behavior of spammer and non-spammer and present an approach that uses call and social network features for making decision about behavior of the subscriber in a service provider network. 2) Spammers normally slowly make calls to recipients of many service providers; however, their behavior remains same across all service providers. For early identification of the spammer, we incorporate collaboration among service providers without having any threat to privacy of subscribers and service providers. 3) Spammers only make few spam calls from one identity and if blocked by the detection system rejoins the network with a new identity. However, spammers have some overlap in the

target recipients of his different identities with more or less similar calling behavior. We incorporated identity linking procedure to link the similar identities that belong to one physical individual and then computed reputation of an individual rather than calling identity. All systems process information from the CDR for the construction of call graph of subscribers and perform certain computations without having interaction with the subscriber. Each system of design approach is described as follows:

**Standalone SPIT Detection:** The Standalone system can be placed in a service provider networks and uses a mechanism that uses subscribers social and call network features for identification of spammers in the network. The standalone system is based on the intuitions that legitimate subscribers normally have long duration calls with many of their called callees thus develop strong social connections with many subscribers and have weak social connections with only few subscribers. On the other hand spammers normally call large number of subscribers which more often results in a large number of small duration calls to many of his callees. This calling behavior thus develops a strong social connection with only few subscribers and has a weak social connection with the large number of subscribers. We incorporated behavioral patterns of subscriber for computing reputation of the subscriber with in the network. The system finally classifies subscriber as a spammer and non-spammer based on reputation scores and automated threshold below which subscriber are flagged as spammer. We called standalone system as Caller-REP (Caller-REPutation) and Chapter 4 provides further details on the approach used within Caller-REP system and its deployment in a real network.

**Collaborative SPIT Detection:** The collaborative system in a proposed system is termed as COSDS (COLlaborative SpIt Detection System). COSDS perform its operation by have privacy-aware collaboration among independent standalone reputation systems deployed in the service provider network. The COSDS system is based on the following observations: 1) spammers distribute low rate spam calls to subscribers of many service providers without overwhelming any single service provider with a high rate spamming. 2) The behavior of spammers remains same across all service providers. Having collaboration among service providers would greatly improve the detection accuracy and decreases detection time, but it has challenge of convincing the service provider to be part of collaboration process. Service providers are not willing to take part in the collaboration because to them collaborating with peer service provider means exchange of information which might be threat to the privacy of their customers and its network configurations. However, use of trusted centralized repository and exchange of non-sensitive filtered information to the centralized repository would somehow convince service provider for taking part in the collaboration process. Chapter 5 provides further details on the design of a privacy-aware collaborative detection system that involve collaboration among service

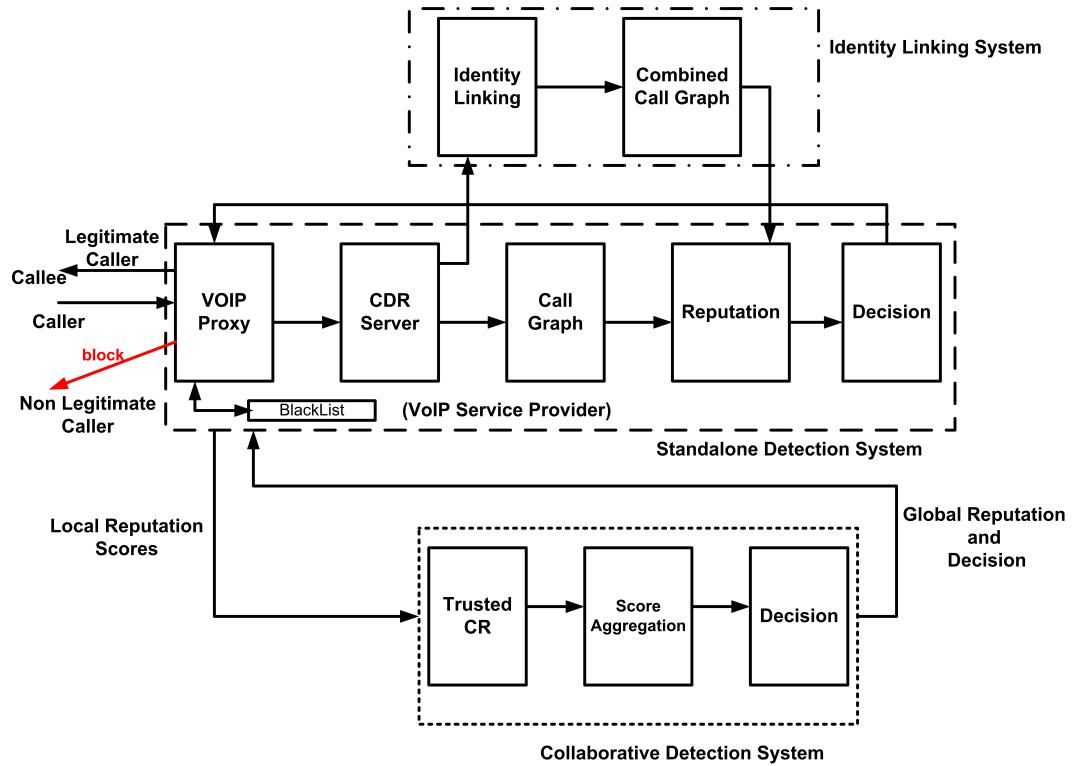


Figure 3.3: Block Diagram of SPIT Detection System.

providers without making any threat to privacy of subscribers of collaborating service provider.

**Identity linking and Spam Detection:** The third component of proposed system is termed as EIS (Early Identification of Spammer) – an identity linking and Spam detection system that uses call and social network features of identities to connect similar identities together and perform spam detection process. The intuition for the design of EIS system is based on the following observations: 1) spammers frequently change their identities in-order to remain undetected and 2) spammer has overlap in a call network among his several identities with more or less similar call patterns. The linking of identities that belong to one physical individual would identify the spammers having multiple identities in a timely way. Chapter 6 provides further details on the design of EIS system and its effect on the identification of the physical spammers having multiple identities.

### 3.3 Call Detail Records

Telecommunication service providers (VoIP, Mobile, and Legacy Telephony) records call transactions of their subscribers in a Call Detail Record (CDR) that are basically used for

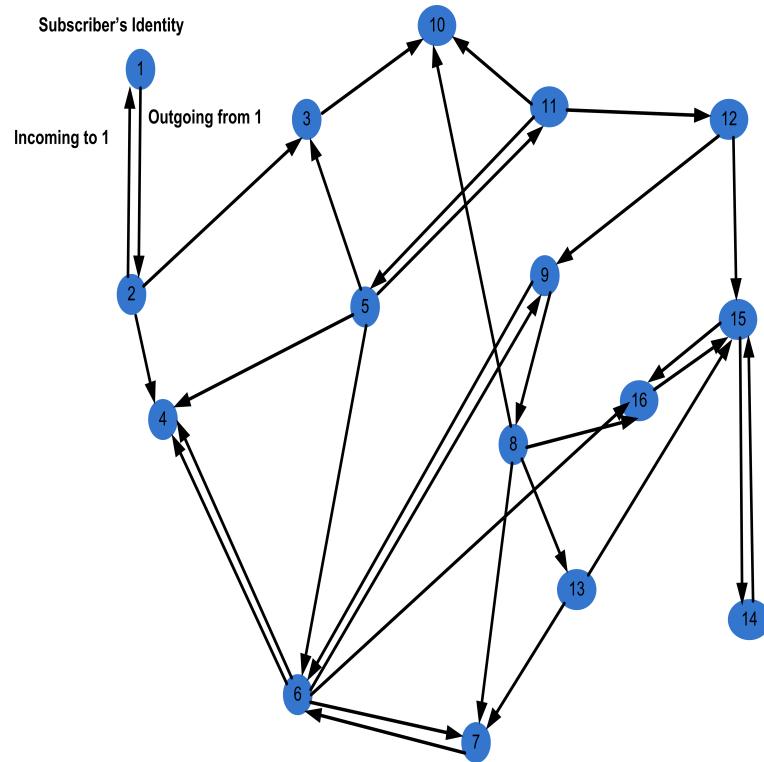


Figure 3.4: Social Network of Subscribers Extracted from the CDRs.

billing purposes and network management. Service providers can utilize these call records for characterizing the behavioral patterns of their subscribers for other purposes such as marketing, personalized offering of new products and identification of malicious users targeting legitimate subscribers. CDR normally contains meta-data of call transactions without any recorded speech contents. A typical CDR widely consists of many fields but of them few are enough for characterizing the behavior of the user. These fields are: identities of a subscribers involved in a call (Caller and the Callee), time of call when subscriber initiates the call to the callee, time when call disconnected by caller or the callee, duration of a call, who disconnected the call, call type (voice, SMS,MMS) and status of the call (successful or failed) etc. In this thesis, we modeled behavior of the user using four fields and construct a weighted social graph of the subscriber. These fields are: calling identity of the caller and callee, time of the call and the call duration.

## 3.4 Social Call Graph

Since CDRs stored detail information about call transactions (incoming and outgoing call) of the subscriber but are normally available in a raw form. The raw call records need to be

processed in order to have meaningful information for business intelligence and identification of malicious subscribers abusing other subscriber or service provider for financial benefits. A complete weighted call graph  $G$  of the subscriber is required for analyzing the behavior of the subscribers towards others. A weighted call graph  $G$  is represented as  $G(N, E, W)$  which is generated for each identity or subscriber present in the raw call records. In  $G(N, E, W)$ ,  $N$  denotes the set of vertices representing the VoIP subscriber which can be either caller or the callee or both,  $E$  denotes the a set of links between subscribers, and  $W$  denotes the weights on the links representing social strength between subscribers. The  $N$  can be either Caller  $S$  or a Callee  $R$  where  $(S, R) \in A$ . Specifically, if  $S$  is the caller and  $R$  is the callee then an edge exists between  $S$  and  $R$  if  $S$  and  $R$  interacted with each other at-least once. Social call graph can be directed or the undirected depending on the type of network. The direction of the link determines whether call is outgoing call from the subscriber or the incoming call to the subscriber. The weights on the links can be assigned from the call- An example call graph of subscribers from the CDRs is shown in a Figure 3.4 and is represented as a sparse adjacency matrix, where 1 represents that caller  $S$  has interacted with callee  $R$  and 0 represents that there happen no interaction between caller and the callee. A sparse adjacency matrix  $A$  of subscriber is represented by an nxn adjacency matrix  $A$  with elements as:

$$A_{ij} = \begin{cases} 1 & \text{if } i \text{ interacted with } j \\ 0 & \text{no interaction} \end{cases} \quad (3.1)$$

In-case of weighted call graph,  $A_{ij}$  are replaced by the weights determined from the frequency of interaction and call duration of interactions. In this thesis, a weighted call graphs is constructed by extracting the following three parameters from the call records for the specific time period.

**Call Duration:** Call Duration represent the time two subscribers spoke to each other. Specifically, out call duration of caller  $S$  to a callee  $R$  is the sum of duration of all calls made by caller  $S$  to callee  $R$  and incoming call duration of caller  $S$  is the sum of duration of calls  $S$  received from callee  $R$ . The aggregated call duration, therefore is the sum of call duration of all calls made and received by the subscriber  $A$ .

**Call-Rate:** Call-Rate represents the frequency of interaction between caller and the callee. Specifically, out call rate between caller  $S$  and callee  $R$  is the sum of all calls made from caller  $S$  to callee  $R$  and in call rate of  $S$  is the sum of calls made by  $R$  to  $S$ . The aggregated call-rate, therefore is the sum of all calls made and received by the user  $A$ .

**Partners:** Partner is the total number of unique subscriber a certain subscriber initiated calls to or received call from and can be grouped into incoming and outgoing partners.

The incoming partners of caller  $S$  is represented as  $P_{IS}$  and out-going partners of caller  $S$  is represented as  $P_{OS}$ . The out-going Interactions represents that user is more important to the certain user than those he did not initiate any call.

The goal in this thesis is to compute the direct trust and global reputation of the subscriber from the weighted call graphs  $G$  and then classifies subscriber as a spammer and a non-spammer.

## 3.5 Social Network Features

People use telephony for the interactive communication with each other and develop weak and strong social relationship with others over the time. Spammers also try to exploit the telephony network for financial intent (e.g. marketing of products, advertising, visiting, frauds etc.) and also develop strong and weak social network with many users. However, the social behavior of legitimate subscriber is different from the social behavior of spammers when analyzed in perspective of different social network and call features. This section presents social and call characteristics of subscriber that can help in differentiating spammers from the non-spammers. We outline calling behavior of spammers and non-spammers for the following features: number of callees the subscriber calls, number of callees calling the subscribers, call duration of subscriber's incoming and out-going calls, incoming and out-going call rate of the subscriber, centrality measure and reciprocity measure of the subscriber.

### 3.5.1 Degree

One of the most important structural measures of a subscriber in a social call graph is the degree of the subscriber. The degree of a subscriber in a social call graph is sum of subscribers he received and made calls. In a directed social call graph, the subscriber can have two degree measures: the out-degree and the in-degree. The out-degree of a subscriber  $i$  in an adjacency matrix  $A$  is the sum of the row entries and the in-degree of subscriber  $i$  is the sum of the column entries associated with the subscriber  $i$  and can be represented as:

$$\text{Out-degree}_i = \sum_{j=1}^n A_{ij} \quad (3.2)$$

$$\text{In-degree}_i = \sum_{j=1}^n A_{ji} \quad (3.3)$$

Where  $A_{ij} = 1$  if there is an outgoing link from subscriber  $i$  to subscriber  $j$ , and zero otherwise. Similarly,  $A_{ji} = 1$  if subscriber  $j$  has out going link to subscriber  $i$  and zero otherwise. In case of a weighted network, the out-degree and in-degree is simply sum of weights of rows and columns of the subscriber  $i$ . The degree of the subscriber can also be represented as the degree distribution which is the probability distribution of the subscriber's degree over the degree of a whole network. Many real networks such as World Wide Web [NEW05a], phone call graphs [NSC+08], [NGD+06], network of autonomous IP systems [FFF99] and online social network [MMG+07] exhibit a power-law degree distribution. It might be possible that inclusion of a large number of spammers in a network would divert the degree distribution from a power-law degree distribution [MOT12].

In a VoIP and voice networks, spammer normally calls large number of subscriber and hardly receives calls from a very few recipients thus has unbalanced out-degree and in-degree structure. On the other hand, legitimate subscriber calls limited number of callees and normally receives calls from many of his callees thus results in a balanced out-degree and in-degree structure. The threshold on an out-degree and in-degree could be useful for blocking spammers [LY07], but using one feature (small in-degree or high-degree as sign of spamming) alone would result in a high false positive rate and small true positive rate. For example, using high in-degree as a sign that subscriber is legitimate would result in a blocking of some legitimate subscriber such as call centers and organizations having small in-degree but have high number of long duration out-going calls. Similarly, using small out-degree as a sign that subscriber is legitimate would block the legitimate subscriber having high out-degree with high duration calls as well has high in-degree with legitimate behavior. It is important that spam detection should not be limited to the degree distribution (in-degree and out-degree) but is also required to consider degree feature along with other call and social features such as call-rate, call duration and centrality.

### 3.5.2 Call-Rate

Call-Rate is the sum of total number of calls made or received by the subscriber. Call-Rate can also be grouped into two types: in-coming call rate and out-going call rate. A high number of calls between two subscribers represent that they are strongly connected with each other. The in-coming and out-going call rate of a caller  $S$  with a callee  $R$  can be computed as:

$$\text{Out - CallRate}(S \rightarrow R) = \sum \text{Calls from } S \text{ to } R \quad (3.4)$$

$$\text{In - CallRate}(S \leftarrow R) = \sum \text{Calls from } R \text{ to } S \quad (3.5)$$

The aggregate out-going and in-coming call rate of the subscriber  $i$  is represented as:

$$\text{Out-CallRate}(A_i) = \sum_{i=1}^n \text{CallRate}_{ij} \quad (3.6)$$

$$\text{In-CallRate}(A_i) = \sum_{j=1}^n \text{CallRate}_{ji} \quad (3.7)$$

Where  $\text{CallRate}_{ij}$  is adjacency matrix of call-rate of subscriber  $i$  to his called callees  $j$ .

A legitimate subscriber normally has repetitive calling behavior with the large number of subscribers (family, friends) and has non-repetitive calling behavior with the few subscribers (strangers). On the other hand spammers or compromised calling identities would like to reach as many subscribers as possible without repeating a target thus develop a non-repetitive network with many subscriber. A large number of target victims and non-repetitive calling behavior with the large number of called victims can be a strong indication that subscriber is spamming. However, using a call rate feature alone would result in a high false positive and small true positive. For example non-legitimate debt collectors make calls to same recipients for payment of debt or a spammer making spam calls to same subscriber to convince on certain offer. Therefore, it is necessary to use the call rate feature along with other call and social network features. For example using ratio of out-degree and out-call rate. Spammers normally have this ratio near to one while legitimate subscribers would have this ratio near to 0.5 or less than 0.5.

### 3.5.3 Call Duration

In telephony, the call duration is the length of duration subscribers talked to each other and can also be grouped into in-coming call duration and out-going call duration. The out-going call duration of subscriber represents how much subscriber trust and want to talk to other subscriber and in-coming call duration of a subscriber represent how much other subscriber trust and talked to the subscriber. The call duration and call-rate together characterize the strength of social ties between the subscribers. The higher the call-rate and call-duration between subscribers, the stronger the social tie exist between subscribers and smaller the call-rate and call duration between subscribers, the weaker the social tie exist between subscribers. The in and out-call duration between caller  $S$  and the callee  $R$  can be represented as:

$$\text{Out-Call-Duration}(S \rightarrow R) = \sum \text{Talk time from } S \text{ to } R \quad (3.8)$$

$$In-Call-Duration(S \leftarrow R) = \sum \text{Talk time from } R \text{ to } S \quad (3.9)$$

The aggregate out and in duration of a caller  $S$  is represented as :

$$Out-Call-Duration(A_i) = \sum_{i=1}^n \text{TalkTime}_{ij} \quad (3.10)$$

$$In-Call-Duration(A_i) = \sum_{j=1}^n \text{TalkTime}_{ji} \quad (3.11)$$

Where  $\text{TalkTime}_{ij}$  is the adjacency matrix of call duration of subscribers. Similarly, the average in and going call duration of the caller is represented as:

$$\text{Avg.In-CallDuration}(A_i) = \frac{\sum_{i=1}^n \text{TalkTime}_{ij}}{\sum_{i=1}^n \text{CallRate}_{ij}} \quad (3.12)$$

$$\text{Avg.In-CallDuration}(A_i) = \frac{\sum_{j=1}^n \text{TalkTime}_{ji}}{\sum_{j=1}^n \text{CallRate}_{ji}} \quad (3.13)$$

Call duration or average call duration is an important feature for estimating the strength of social relationship between subscribers and is also useful for characterizing the behavior of the subscriber in a network. Legitimate subscribers normally have some good number of long duration calls with his friends, family members and colleagues, and have relatively small duration calls with only few callees for example strangers. On the other hand call recipients are not comfortable talking with unknown subscriber for the long time periods thus disconnect call as soon as they realized the true identity and motivation of the caller. Because of this behavior, spammers normally have large number of short duration calls with their recipients with only small number of long duration calls. However, small duration is not the only sign caller is spammer for example a school announces a short duration urgent announcement to a large number of student. The use of call duration along with call-rate and out-degree of subscriber would provide enough evidence to classify a subscriber as a spammer and a non-spammer.

### 3.5.4 Eigen Centrality

Eigenvector centrality measures centrality of a subscriber in a call graph by computing eigenvector of the largest positive eigenvalue. Eigenvector provides information about

connectivity of subscriber with the other subscribers. Subscribers connected to high reputed subscriber or more significant subscribers would have high reputation score than those subscribers connected to the non-reputed subscribers. The Eigen centrality measures of the subscriber from an adjacency matrix can be computed as:

$$EC(A) = \sum A_{ij} * x_j \quad (3.14)$$

Where  $A_{ij}$  is the adjacency matrix of a graph  $G$  and  $x_j$  is the initial centrality score of subscriber  $j$ . Spammers normally choose large number of reputed recipients, thus probably would have high centrality score, whereas legitimate user only have connections with few reputed users thus results in a small centrality score than the spammers. Considering, high centrality score as a sign of legitimacy would result in blocking many legitimate subscriber and allowing of many spammers. However, we believe that Eigen centrality would not only be limited to structure of the network but also need to consider the relationship strength between subscribers. Computing centrality by considering the trust weights would probably results in a small centrality score for the spammers and high centrality score for the non-spammer.

### 3.5.5 Reciprocity

Reciprocity is defined as the fractions of edges that are reciprocated i.e. two subscribers receive and make call to each other. Reciprocity measures the relationship strength between subscribers and is defined as:

$$Reciprocity_i = \frac{OD_i \cap ID_i}{OD_i} \quad (3.15)$$

Where  $OD_i$  is the set of subscriber that are called by the subscriber  $i$  and  $ID_i$  is the set of subscribers that calls the subscriber  $i$ . Reciprocal connection normally exists between the people having low degree and high connectivity [HS08]. Reciprocity measure can be used to identify spammers in a network because legitimate and spammer exhibit different reciprocity measures. Legitimate subscriber tends to have calls in both directions that characterize the strong social relationships of subscriber with his reciprocated subscribers, thus would have high reciprocity score. On the other hand, spammer makes calls to a large number of recipients and many of target recipients are interested in calling back the spammer thus spammer developed a one way weak network with many of his called recipients. This calling behavior and structural imbalances of spammer between incoming and out-going calls from the same recipients would results in assignment of a small reciprocity score to the spammer.

### **3.6 Discussion**

The service provider record all call transactions among its subscribers in the call detailed records which is mainly used for billing and network management such as QoS measurements, identification of spammers and fault detection etc. Social call graph between subscribers of the network can be easily constructed from identity of subscriber and can be used to infer the behavior of the subscriber in the network. The subscriber can be grouped into two types: the legitimate user – using network resources according to signed agreement and the malicious or non-legitimate users – using network resources for malicious activities i.e. making spam calls or trying to steal private information or money from the subscriber.

The legitimate subscribers and the spammer exhibit different social behavior that could be used for classifying them as spammer and the non-spammer. A number of social network features have been proposed that differentiates spammers from non-spammer based on their degree of their unique interactions, reciprocity measures, clustering coefficient etc. [LY07], [WA10], [AMF10], [BR05] . However, in telephony network, two important features i.e. frequency of interactions and call duration of interactions needs to be considered along with structural properties of call graph for estimating the behavior of the subscriber within the network. The combine use of different social network and call feature would possibly improve the detection time, true positive rate and minimize false positive rates.

To achieve objectives of effective detection and minimize detection time this thesis proposes the combine use of social and call features of the user for modeling the behavior of the subscriber used for differentiating spammers from the non-spammers. The contributions of this thesis are three folds: 1) a standalone detection system-that considers the collective use of call duration, call rate and out-degree of subscriber for the computation of reputation of the subscriber in the network, 2) a collaborative detection system-that considers privacy-aware collaboration among standalone detection systems deployed with the service providers in-order to minimize detection time and identifies smart spammers, and 3) an identity linking system-that connects multiple identities that belong to one physical person and then collectively compute reputation of the physical person considering all his identities together.

# **Chapter 4**

## **Caller-REP: Detecting Unwanted Calls Through Caller's Social Strength**

### **4.1 Introduction**

The affordable calling rates of VoIP telephony to any destination across the world, and its easy integration with the Internet technologies for the voice communication has attracted a large number of subscribers for business and personal communications. The benefits of VoIP include low management and maintenance costs, scalability, flexibility, and cost savings of a using one network for the voice, data and video. These benefits have also attracted large number of spammers and telemarketers for the unsolicited communications. VoIP spam not only brings financial loss to the subscribers of the technology but also causes displeasure among call recipients with the real-time ringing alerts. VoIP spam is much more intrusive than the email spam as call recipient has to respond immediately and it is too late if detected after the call has already been established.

A voice call consists of two phases - a call setup phase followed by the speech exchange phase. The content-based anti-SPIT systems could be effective for detecting spammers but is already late as spammers have already disturbed callee with the ringing alerts. Moreover, processing speech contents in a real-time requires sophisticated system resources and would certainly add delay in conversations. The ideal SPIT detection system is required to block spammers without content processing and not involving subscribers for the additional feedback about the received calls. The existing identity-based reputation systems use only one feature for modeling the behavior of subscribers that is further used for classifying subscribers as a spammers and a non-spammers. These systems can be easily evaded by spammers by circumventing only one feature. Service provider wishes to have a spam filtering system that fulfills four requirements: 1) the de-

tection system must not require any interaction with the subscriber at any stage of call processing, 2) the system collectively uses number of social and call features that would be difficult for the spammers to circumvent, 3) it automatically computes the threshold below which the subscriber is placed in the blacklist, and 4) it does not require any change in the call setup messages and handset of the subscribers.

This chapter presents Caller-REP (Caller-REPutation), a new approach for ranking VoIP subscribers and then blacklisting SPIT caller without any content processing, user involvement, and changes in the VoIP network infrastructure. The main procedure starts by creating the social call graph by using information from the past call transaction of subscribers recorded in the call detailed records (CDR). Our approach is based on two observations. 1) Legitimate subscribers tend to call the same recipients several times, receive call back from the called recipients, and make long duration calls [SMS+08], [BSG+11], [WBS+09] to a relatively large number of called recipients. 2) On the other hand telemarketers and advertisers tend to have the opposite calling behavior: they try to call a large number of people to deliver their messages [SMS+08], [DTN11] that often result in a short duration calls to the large number of called callees. These subscribers also have fewer incoming calls than outgoing calls, as large number of subscribers are not interested in calling back to telemarketers.

For each VoIP subscriber in a network, Caller-REP first computes direct trust between subscriber and his called callees. The direct trust between subscriber and his called callee is computed from the following social and call features: number of calls made and received between subscriber and the callee, call duration of received and made calls between subscriber and the callee, and the number of unique callees of a subscriber. The global reputation of a subscriber across the network is then computed through power iteration method using normalized direct trust scores of a subscriber. Finally, Caller-REP computes classification threshold by applying 25th percentile based approach to the global reputation scores of the subscribers. The subscriber having reputation score less than threshold are barred from calling. Service providers can also adopt other threshold according to their spam detection policies and tolerance against spammers. Despite having a non-intrusive reputation-based SPIT detection system, another major challenge is protecting privacy of subscribers while computing trust and reputation scores. This chapter also identifies method for using the information from the CDRs in a privacy-aware manner so that intruders at reputation engine would not be able to learn relationship network of the subscribers.

This chapter also analyzes the performance of Caller-REP using the synthetic data set that has been generated by modeling the social behavior of spammers and non-spammers for different social network features. Our results show that Caller-REP is able to block

all spammers within three days. Caller-REP achieves a false positive rate of less than 10% and true positive rate of almost 80% in the first two days even in the presence of a significant number of spammers. This true positive rate would further improve to 99% and false positive rate drops to less than 2% in three days. In a network with no spammers, Caller-REP achieves a false positive rate of less than 10% and in a heavily saturated network with more than 60% of spam callers; it achieves a true positive rate of 98% and no false positives. We also compare the performance of Caller-REP with a closely related spam detection approach named Call-Rank which shows that the Caller-REP system outperforms Call-Rank in terms of detection accuracy and detection time without involving subscriber for the feedback.

This chapter makes the following contributions:

- We present a social reputation-based SPIT detection system that effectively characterizes social behavior of the subscribers using set of social network and call features. To this extent, we modeled behavior of the subscriber as a weighted call graph, compute direct trust of subscriber with his called callees and finally compute reputation of the subscriber using power iteration algorithm. The collective use of social and call features ensures that spammers would not easily circumvent the detection system and would always have small reputation scores.
- We present a percentile based approach for computing automated threshold below which the subscriber is classified as spammer. To this extent we adopted a 25th percentile based approach which is also tunable according to the requirement of service provider.
- We present a privacy protection mechanism that ensures that privacy of subscriber is not breached during computation of reputation scores. The privacy of subscriber is protected by exchanging pseudonymized information between reputation engine and the call processing engine.
- We evaluated the proposed system on a synthetic data set that has been generated by simulating the social behavior of the spammers and non-spammers. Moreover, we have analyzed the performance of system under different conditions such as high spamming rate, small spamming rate, and for different performance metrics.

The rest of this chapter is organized as follows. Section 4.3 describes the design of Caller-REP system and algorithms used for the computation of direct trust, global reputation and automated threshold. Section 4.4 provides simulation setup that has been used for generating the synthetic data-set. Section 4.5 evaluates the Caller-REP's performances for different performance metrics using different percentages of spammers and non-spammers.

Section 6.5 briefly discusses important features of Caller-REP system and possible evasion approaches, and Section 4.7 concludes the chapter.

## 4.2 Motivation

The challenge in a design of SPIT detection system is to block a SPIT caller during the call setup phase – thus without content analysis and user involvement. The existing reputation-based approaches are not only intrusive to the end-users but also use a single feature for computing direct trust and reputation of the subscriber. Spammers can easily manipulate single feature and circumvent detection system for relative long time periods. The average call duration between subscribers can be the sign that they trust each other but trust between subscribers in telephony should not be limited to the average call duration only. A SPIT caller always targets a large number of subscribers and may manage to have good duration calls with many of his target subscribers. This would probably result in a high trust and reputation for the SPIT caller across the network because of some high duration calls to his callees. For example, consider a VoIP network with  $n$  subscribers. A SPIT caller makes calls to a 50% of total subscribers and manages to have call duration of 60 seconds to more than 20% of the subscribers it calls. In this case, the SPIT caller would have strong calling network with at-least 20% of subscribers that would result in an aggregate high reputation for the SPIT if average call duration is used as a sign of trustworthiness . This happens because SPIT caller calls a certain subscribers only once thus his aggregate call duration and average call duration remains same over the time. Though the SPIT caller exhibits non-repetitive calling behavior but would manage to circumvent the average call duration-based reputation systems.

In voice networks, the legitimate subscribers normally have a repetitive calling behavior with his friends and family members, whereas the spammers normally do not have repetitive calling behavior. Legitimate subscribers also receive calls from friends and family members thus have mutuality in their social network, whereas the spammers do not receive good duration calls from a large number of their target victims. This behavior would result in an unbalanced calling network of spammers (few in-coming calls and large number of out-going calls). We believe that trust and reputation of the subscriber should be computed collectively using call duration of subscribers in both directions, call-rate of subscribers in both directions and out-degree of the subscribers. The collective use of these features for computing trust and reputation of the subscriber would result in a small reputation score for the spammers, despite small number of high duration calls with their victims.

## 4.3 Caller-REP:Caller Classification-Based on Reputation

This section describes requirements for the reputation-based SPIT detection system and data source used for non-intrusive SPIT detection system. Moreover, this section also describes components and algorithms of a Caller-REP system. Finally, we have also compared Caller-REP system with other reputation-based systems.

### 4.3.1 Requirements for Reputation Based SPIT Detection System

Before describing our reputation-based SPIT detection system, it is important to point out few requirements for the design of an effective SPIT detection system based on the trust and reputation of subscribers.

- The computation of direct trust between subscribers and the global reputation of the subscriber must not involve subscribers for the feedback at any stage of call processing thus must have ideally zero subscriber's involvement.
- The computation of the reputation and the direct trust of the subscriber needs to consider the subscriber's past call transactions within the network along with the collective use of social and call features.
- The proposed system must be robust against different malicious attacks such as Sybil attack and must not be easily circumvented by the spammers.
- The system must carry out all computation in anonymized way in order to protect the privacy of network subscriber.
- The overall system must not require any changes in the network or handset of the subscriber.
- The system must be tune-able in terms of classification threshold according to requirement of service providers or carriers and can be easily integrated with other systems.

### 4.3.2 Data Source

A VoIP service provider consist of a large number of networking devices that work together for providing telephony and messaging services to its subscribers. Each call transaction either received or made by the subscribers is recorded in the call logs at the proxy servers or call handling engine, that are later pushed to the CDR server for the billing and troubleshooting. The logs contain several fields providing information about the call

transaction. The time-stamp provides information when call between subscribers established , call duration represent the how much time two subscribers talk to each other,, caller id reflects the calling identity of subscribers. Besides these fields, the CDR may also logs IP address of caller and the callee, type of call (SMS, voice or MMS), who disconnected call etc.

In order to compute global reputation of the subscriber, we construct a directed call graph between subscribers by extracting the following information from the CDR logs.

**Call Duration:** Talk time of the subscriber with his called callee. The subscriber can be either caller or the callee. We are mainly interested in characterizing behavior of subscriber as a caller. We represent call duration between the caller  $S$  and the Callee  $R$  as  $CD_{SR}$  which is the sum of duration of all calls made from a caller  $S$  to the callee  $R$ .

**Call Rate:** Call-Rate is the number of calls made and received by the caller. We present call rate between caller and the callee as  $CallRate_{SR}$  and is the sum of calls made from the caller  $S$  to a callee  $R$ .

**Partners:** Partner is the number of unique callees a caller has interacted or received calls from. We represent number of out partners as a  $PO_S$  which is the sum of unique callee a caller  $S$  is calling.

Our goal is to use information from directed call graph and develop a SPIT detection that classifies subscriber as a spammer and a non-spammer based on reputation of the subscriber. The block diagram of a Caller-REP is shown in a Figure 4.1. Specifically, after modeling the call graph from the CDRs, Caller-REP consists of three steps for classifying subscriber as a spammer and a non-spammer: 1) it computes direct trust of the subscriber with his called callee using social network and call features, 2) It computes global reputation of the subscriber using power iteration algorithm, and 3) It automatically computes the threshold below which subscriber is classified as spammer.

### 4.3.3 Subscriber Direct Trust

The calling behavior of the subscriber can be estimated from the level of a trust a subscriber maintains with other subscribers. Trust represents the level of mutual relationship between subscribers which they have developed over the time and is computed from their direct call transactions. Direct Trust between subscribers represents the amount of duration and number of times the subscribers interacted to each other. Higher the call duration and call rate between subscribers, higher the trust exists between them. This trust information is then extended to estimate the network wide behavior of a subscriber termed as reputation of the subscriber.

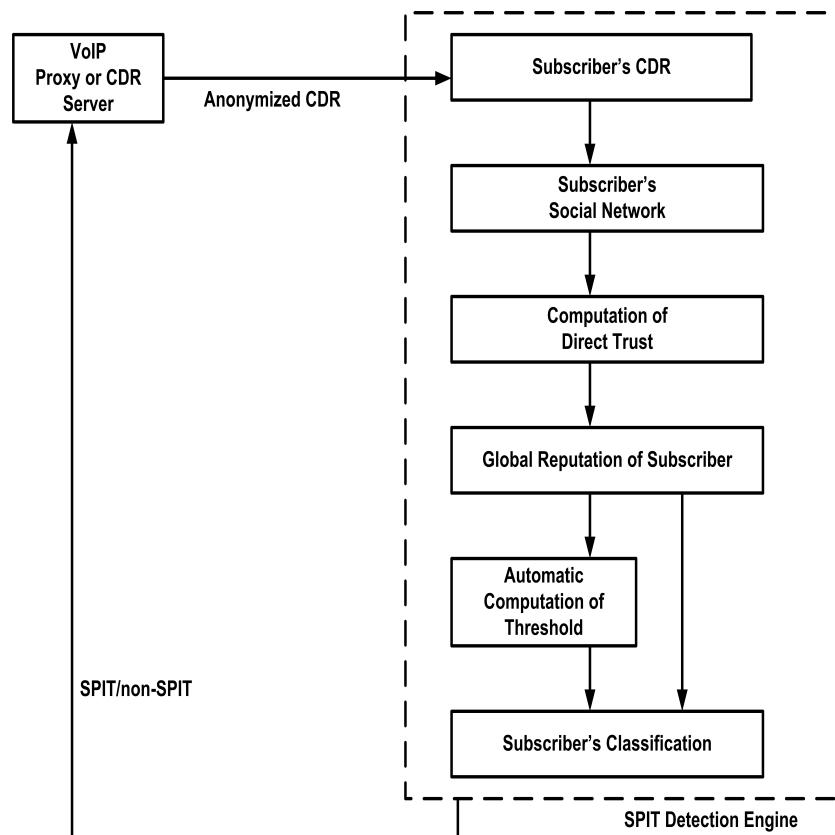


Figure 4.1: Building Blocks of Caller-REP.

The direct trust between subscribers characterizes the strength [GRA73] of social relationship between subscribers. In a voice network, subscribers can develop strong relationships if they have repetitive and reciprocated long duration calls to each other; and can develop weak relationship if subscribers have non-repetitive and non-reciprocated short duration calls to each other. In perspective of social behavior of legitimate and spam subscribers, we argue that legitimate subscribers usually have strong social ties with a large number of callees and weak social ties with a few callees , whereas a SPIT caller develops a weak relationship with a large number of their called callees. This trust information or relationship information can be used to classify subscriber as a spammer and a legitimate.

In existing trust and reputation based SPIT detection systems, the direct trust between subscribers is computed in two ways: getting positive or negative feedback from the subscribers about their callers from their past call transactions and 2) implicitly using information from the call detail records. First approach is intrusive to subscribers and annoys subscribers for the feedback. Moreover, it also require changes in the handset and call setup messages which makes it infeasible to be deployed in a real VoIP network. Second approach normally uses single feature for computing direct trust between subscribers i.e. average call duration, but we argued that direct trust should not be limited to the single

feature but should also consider number of features for computing direct trust between subscribers.

As spammers normally target large number of callees and managed to have good duration calls with many of their callees. The use of average call duration feature for trust computation would result in a high trust score of the spammer with many of his callees. This in turn would have high reputation score despite having a large number of out-going calls and small number of incoming calls. For example, a subscriber with sum of call duration of 10 minutes in 10 calls has average call duration of 1 minute which is similar to the subscriber having sum duration of 1 minute in 1 call because average duration is 1 minute. Additionally, high number of long duration calls would also make spammer as a legitimate subscriber because a group of people may also be interested in spam call because of greed for financial benefits or the calls that terminated on the voice mail box. In a voice communication network, a subscriber can be either a caller or the callee. In the rest of this chapter we are interested in dealing subscriber as a caller. The subscriber's transactions are represented as a sparse adjacency matrix in which rows represent callers and column represents the callees. We are interested in computing the direct trust between caller and the callee which is also represented as a sparse matrix. We incorporated the following features for computing the direct trust between caller and the callee: the frequency of interaction between caller and callee in both directions (incoming and outgoing), call duration between caller and the callee in both directions (incoming and outgoing) and the out-degree of the caller (number of unique callees of the caller). The call duration ( $CD_{SR}$ ), call-rate ( $CallRate_{SR}$ ) and out-degree vector ( $PO_S$ ) are collectively used to estimate the direct trust between subscriber  $S$  and his callee  $R$  using equation 4.1.

$$Trust_{SR} = \frac{CD_{SR} \times CallRate_{SR} + CD_{RS} \times CallRate_{RS}}{PO_S} \quad (4.1)$$

For all subscribers in the network, the direct trust matrix is defined as  $N \times N$  sparse matrix. In equation 4.1,  $Trust_{SR}$  represents the trust score of subscriber  $S$  with his callee  $R$  based on their direct call transactions. In equation 4.1  $CD$  is the call duration between subscriber  $S$  with his callee  $R$ ,  $CallRate$  is the interaction rate between subscriber  $S$  with his callee  $R$ , and  $PO$  is the out-degree of the subscriber  $S$ . The  $Trust_{SR}$  is represented as  $N \times N$  adjacency matrix where row represent caller and column represent callee. The direct trust computed from equation 4.1 would result in a small direct trust scores for subscribers having large number of out-going partners with large number of small duration calls and have small number of incoming calls. On the basis of high call duration, repetitive calling behavior and small out-degree, equation 4.1 would assign a high trust

score to the legitimate subscriber and small trust score to spammers due to their high out-degree and non-repetitive short duration calls. The spammer would only be able to achieve high direct trust scores if he exhibits following behavior: 1) managed to have large duration repetitive incoming and outgoing calls, and 2) develop strong relationship with many callees.

The reputation of the subscriber is computed from the normalized direct scores and present network wide view about behavior of the subscriber. The normalized direct trust matrix is computed by dividing each element of a row by summation of the respective row as shown in equation 4.2. This ensures that all trust scores will be between 0 and 1 and each row would sum to 1 as  $\sum_R T_{SR} = 1$

$$T_{SR} = \frac{Trust_{SR}}{\sum_R Trust_{SR}} \quad (4.2)$$

The direct trust scores provide information about direct relationship between subscribers and would characterize how strong or weak relationship exists between subscribers. Once the direct trust score of the subscriber has been computed the next step is to aggregate these normalized direct trust score to have network wide view about behavior of the subscriber.

#### 4.3.4 Reputation of the Subscriber

The global reputation represents the aggregate behavior of subscriber towards his entire interacted subscribers across the network. If a particular subscriber has no prior interaction with the other particular subscriber then the subscriber would ask other subscribers for the feedback about the subscriber. In this situation, the global reputation of the subscriber within the network would provide information about the aggregate behavior of the subscriber towards all his interacted subscribers. In this section, we outline procedure used for computing the global reputation of the subscriber.

The global reputation of the subscriber is computed by aggregating the normalized direct trust scores. Eigen Trust algorithm [KSM03] has been widely applied in P2P network for computing reputation of the node from his direct trust scores. In Caller-REP, the reputation of the subscriber is computed using power iteration method with a slightly different initial reputation scores. The input to reputation computation method is the matrix of normalized direct trust  $T_{SR}$  between each pair of subscribers ( $S, R$ ). The output of this algorithm is a global reputation vector  $G$  with a global reputation score of a subscriber  $G_S$  and is in between  $\in [0, 1]$ . The algorithm first initializes the initial global reputation values of each subscriber with the inverse of the out-degree  $PO_S$  of the subscriber  $S$ . The

global reputation score of subscriber is then iteratively computed by multiplying the normalized direct trust matrix and initial reputation vector as represented in equation 4.3 and algorithm 1. The iteration process stops on the convergence of the norm of the global reputation vector  $\|GR\| = \sqrt{\sum_S GR_S^2}$ . In each step of this iteration process, the global reputation vector is updated from the normalized direct trust and global reputation score as  $GR = T \times GR$ .  $GR$  is normalized and its norm  $gr$  is used along with the previous norm  $gr_{previous}$  for checking the convergence.

$$GR(t+1) = Trust_{SR} * GR(t) \quad (4.3)$$

The reputation computed in this way would result in a high reputation score for the subscribers having long duration repetitive calling behavior with the reputed subscribers. The spammers in this case would have a small reputation scores because of unbalanced calling behavior and large number of recipients.

---

**Algorithm 4.1** Reputation Computation

---

```

1: procedure GLOBAL REPUTATION OF ALL SUBSCRIBERS {S}
2:   input  $\leftarrow$  Trust (normalized direct trust matrix, with elements  $T_{SR}$ )
3:   output  $\leftarrow$   $GR$  (Global reputation score vector, with elements  $GR_S$ )
4:   precision parameter  $\leftarrow \epsilon$ 
5:   Initialize reputation vector  $GR$ 
6:    $GR_S = [1/PO_S]$ 
7:   % Iterate until convergence
8:   while  $\delta < \epsilon$  do
9:      $GR \leftarrow Trust \times GR$ 
10:     $GR \leftarrow GR / \|GR\|$ 
11:     $gr \leftarrow \|GR\|$ 
12:     $\delta \leftarrow \frac{gr - gr_{previous}}{gr}$ 
13:     $gr_{previous} \leftarrow gr$ 
14:   end while
15: end procedure

```

---

### 4.3.5 Detection of Spammers

The reputation of a SPIT caller deviates from the reputation of legitimate subscribers and this deviation would help in distinguishing SPIT subscribers from the non-SPIT subscribers. In this section, we provide a method used for computing the automated threshold below which the subscribers is classified as a spammer.

The global reputation score of the subscriber can be used in three ways to decide about the nature of the subscriber. 1) The global reputation scores can be sent to the callee as a

part of a SIP invite message and callee decides whether to accept or reject the call. 2) The global reputation scores can be compared with a fixed threshold. 3) The global reputation scores can be compared with a dynamic threshold learn from the set of reputation values. The first approach requires interaction with the callee thus is not only intrusive but also requires changes in the call setup messages. In the second approach, the threshold is decided based on a certain pre-defined true or false positive rate but this threshold is not flexible to accommodate a continuous changing behavior of the subscriber. In the third option, the threshold is computed automatically from the reputation score without subscriber's intervention and can also account the changing behavior of the subscriber. We adopted third option for computing the automated threshold below which the subscriber is classified as a spammer. The dynamic threshold is advantageous as compared to the fixed threshold approach as it better minimizes false positive rate and maximizes true positive rate by considering the traffic patterns of subscribers within a specific time window.

We set the dynamic threshold value based on a percentile method instead of a fixed value threshold. We sort the set of computed global reputation scores of all subscribers and set the threshold value to the 25th percentile of this set. The procedure for classifying subscriber as a SPIT or a non-SPIT is presented in algorithm 2. As in algorithm 2,  $GR$  is the global reputation vector of all subscriber and  $m$  is the 25th percentile value of the global reputation. First, the 25th percentile of global reputation is computed for each time window. Second, the mean of global reputation score of subscriber less than the 25th percentile value  $m$  is set as a dynamic threshold for a specific time window.

---

**Algorithm 4.2** Detection of Spammers

---

```

1: procedure SPIT DETECTION
2:   input  $\leftarrow$  Global Reputation( $GR$ ), with elements  $GR_S$ 
3:   output  $\leftarrow$  SPIT (1) or non-SPIT(-1) detection vector, with elements  $SPIT_S$ 
4:   service provider-defined parameter  $\leftarrow \beta$  ( $\beta = 1$  if service provider has no preference)
5:    $m \leftarrow$  1st-quartile( $GR$ )
6:   threshold  $\leftarrow$  mean( $GR < m$ )
7:   for All subscriber  $S$  do
8:     if ( $GR[S] < \beta \times \text{threshold}$ ) then
9:       Place Subscriber  $S$  in a SPIT List
10:      else
11:        Do Not Place Subscriber  $S$  in a SPIT List
12:      end if
13:    end for
14:  end procedure

```

---

Subscribers can be classified as legitimate 1 or non-legitimate -1 based on a following rule:

$$\text{Subscribers}_S = \begin{cases} GR_S > \beta \times \text{threshold} & ; 1 \\ GR_S < \beta \times \text{threshold} & ; -1 \end{cases}$$

Other approaches like Inter-quartile distance, mean absolute deviation and machine learning-based approaches can also be used for identifying the suspected SPIT subscribers. We believe that at any given time period, the VoIP network possibly has SPIT traffic less than 25% of the total incoming traffic and a dynamic threshold based on the 25th percentile would achieve better true positive rate. The service provider also wishes to block all the top spammers. Moreover, technologies normally witness a few malicious users until they become mature and attract large number of users. However, once it attracted large number of users then, it also starts attracting large number of malicious users and the percentage of spammers rises to as up as 40% of all identities joining the network on a particular day. For example, currently, it is estimated that 36% tweets on tweeter contains links [TWI16] and 25% of all personal computers may be infected by viruses. Over the time, we expect similar behavior in case when VoIP and telephony becomes affordable and a primary method for having personal and business communication. Furthermore, this threshold is tunable and can easily be integrated with service provider policies against spammer and non-spammer. However, the 25th percentile based approach would not provide good results if the percentage of SPIT traffic increases or if detection window size is decreased for the reputation computation. The 25th percentile performs better when large window size is used for the computation of reputation scores i.e. large number of call records and user interactions. The true positive and true negative rates in a network with high SPIT or legitimate traffic would be maximized by using Caller-REP with a  $\beta$  parameter set by the service provider according to his SPIT detection policies.

#### 4.3.6 Caller-REP System Components

Figure 4.2 shows the Caller-REP and its interactions with call processing system i.e. VoIP proxy server. The Caller-REP system can be implemented in two modes; as a standalone system having dedicated hardware resources or resides on the CDR or proxy server as a detection module. The later implementation would probably increase load on a proxy server and the former requires communication link between CDR server and Caller-REP server. However, in both implementations modes the system would not add any additional delay to the call setup message as the reputation is computed in the background and blacklist is consulted seamlessly during the process of getting billing and authorization

information. The Caller-REP system integrates with proxy server and operates in the following way.

1. On receiving a call request from the subscriber for the particular subscriber residing either in his network or other network, the proxy server first checks whether subscriber behavior is legitimate or not by checking subscriber status in black-list and white-list database. If the subscriber is found in the black-list, the proxy server immediately blocks the subscriber from calling and disconnects his call. If the subscriber is found in the white-list, the proxy server allows subscriber to reach the called subscriber and waits for the call termination.
2. Once a call between subscriber has been terminated, the proxy server records log of the call transaction in a CDR and periodically send anonymized CDR to the Caller-REP system for computing reputation of the subscriber.
3. Caller-REP system on receiving CDRs, extracts call and social network features of subscriber from the anonymized CDRs and computes direct trust and global reputation of the subscriber. The automated threshold is then computed from the global reputation scores that classifies the subscriber as spammer and non-spammer. Finally, Caller-REP engine responds to the proxy server with the results about the subscriber i.e status and global reputation score.

#### 4.3.7 Caller-REP and Privacy

Caller-REP system is based on the use of information from the call detail records to determine if a subscriber is a spammer or a non-spammer. The CDRs contain private information about when, where, and who the subscriber calls along with the length of phone calls and other personal information like billing addresses and IP addresses. The availability of this private information can raise serious privacy concerns and enable illicit activities that can put the subscriber at risk. The telecom service providers have to protect the privacy of their subscriber if they want to use this information for specific purposes - e.g. in our case for the spam detection. The following privacy protection requirements apply. 1) Subscribers have to be informed if their call detail records are being used for any specific purpose like intrusion detection or spam detection [OZ04] and be provided with an opt-out option. 2) The service provider has to keep the subscriber data secure and protected from un-authorized access [OZ04].3) The service provider has to make an effort to hide information that can directly identify the user [PEN04] (e.g. user name and phone number) even when providing authorized access to subscriber data to third parties and

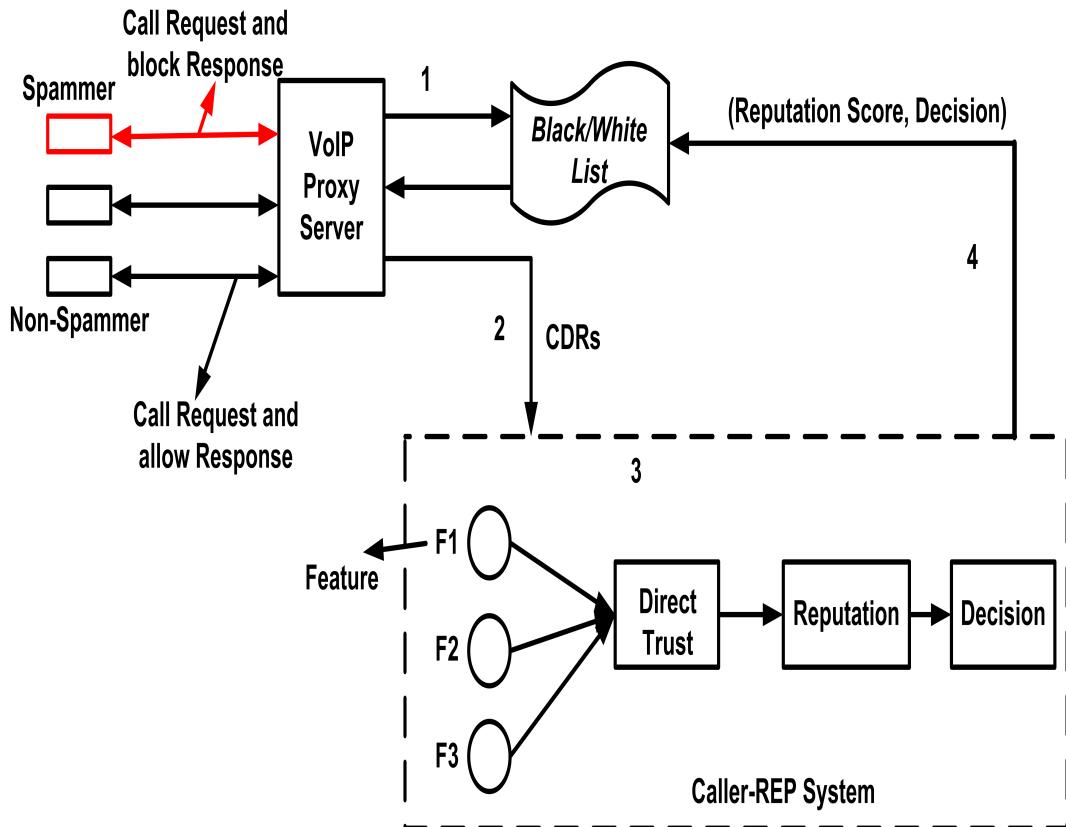


Figure 4.2: Interaction Between Caller-REP and Proxy-Server.

to specific-purposed managed software like spam detection system. This could be done by removing user name and phone number from the data and tagging it with the random identifiers that are unique for each subscriber. This approach does not guarantee that the identity of the subscriber cannot be mapped to the data but makes it non-trivial.

Caller-REP protects the privacy of the subscribers and manages privacy risks in a service provider network in the following way. 1) The service provider shall inform subscriber that their call records will be analyzed for the purpose of blocking spammers and allow subscribers to opt-out only if number of their unique callees are less than some fixed threshold (a small possible threshold, for example 5 unique callees). 2) The service provider should protect user data from unauthorized access using strong authentication process and policies on unauthorized disclosure of subscriber information by their staff. 3) In order to minimize the risk of misuse of data for other purposes, Caller-REP engine also assigns with a data after pseudonymizing [CKK05]. For that purpose, the service provider first selects few attributes from subscriber call transactions (i.e. caller and callee identity, call duration, and time of call transaction). It then replaces caller and callee identities with a key by which he can later re-identify a user account in case that have been positively identified as spammer. Additionally, in order to completely preserve the

privacy of the subscriber we also anonymized the time information and duration fields of the CDR. We anonymized the time information by stripping the seconds and minutes information from the call-time before presenting data to the Caller-REP engine. This is not an anonymization because the service provider is still capable to re-identify the user account. With this design option, the service provider achieves privacy protection requirement 3. That is, the privacy of a subscriber  $S$  who has communicated with callee  $R$  is not compromised because the Caller-REP learns nothing about the subscriber's real identity, thus unable to profile and describe the calling behavior of  $S$  and any other subscriber. Additionally, the intruders having some background information about the particular subscriber would not be able to find the pseudonymized identity of the target subscriber. The drawback of pseudonymization might be the possibility of getting pseudonymized identity of subscriber via data from public information, which cannot be prevented easily. We would provide further detail on privacy preservation for different call features in chapter 5.

#### 4.3.8 Comparison with Other Reputation-based SPIT Detection Systems

There are some reputation-based SPIT detection systems that share some similarity with the Caller-REP system. The major difference between Caller-REP and other reputation-based systems is the method adopted for computing direct trust between subscribers. Additionally, Caller-REP uses slightly different procedure for the initialization of global reputation scores. In this section, we compare Caller-REP system with the closely related Call-Rank and VSD systems. Both Call-Rank and Call-REP computes reputation of subscriber automatically from the CDRs but their procedure is different from each other. On the other hand, VSD computes direct trust between subscribers by getting feedback from the subscriber receiving calls from the particular subscribers.

Call-Rank uses average call duration between subscribers as a feature for computing direct trust between subscribers and then uses Eigen Trust algorithm for computing reputation of the subscriber by aggregating direct trust score. In Call-Rank, subscribers are considered trusted and well reputed if they have high average call duration regardless of their un-expected high number of called recipients. The use of average call duration feature would allow spammers to bypass the system with the little effort. Call-Rank would assign high trust score to spammers, if spammers managed the followings: 1) convinced subscribers even a few to call back them by offering some monetary incentive or prizes. This would increase incoming call traffic and duration to the spammers which in-turn

increase trust and reputation of the spammers. 2) Spammers calls terminated on the recipient voice mail box would result in a high duration calls with many recipients. 3) Spammers usually target large number of recipients only once and managed some good duration calls to few callees. In this scenario, the average duration stays the same as of call duration of one call. 4) The aggregate large number of good duration calls would result in a good reputation score despite having high number of recipients. For example, the average call duration of a subscriber having average call duration of 120 seconds for 10 calls is similar to average call duration of subscriber having call duration of 120 seconds for 1 call. This would consequently results in a same trust score for both subscribers. In all these conditions Call-Rank will assign high reputation scores to both legitimate and malicious subscriber having good duration calls. In perspective of deployment, Call-Rank requires the followings: 1) a private and public key infrastructure for authorization and authentication of subscriber, 2) a mechanism for transporting reputation scores and social network credentials to the callee for the final decision, and 3) a deployment of CAPTCHA or Turing test system for introducing new subscriber and minimizing false positives. The exchange of social network credentials to called subscriber may also be a threat to privacy of calling subscriber as callee is being informed which friends of callee have also been interacted with the caller. These limitations make Call-Rank unlikely to be deployed in a real network.

The major limitation of VSD [DK05], [KD07] is that it requires interaction with the callee for the feedback about the subscriber. The system requires honest feedback and interaction between different system stages for high detection accuracy. However, having honest feedback from all subscribers is not always happening in a real network. Spammers can easily manipulate feedback scores by creating a Sybil network or a network between their own identities. Additionally, subscribers get annoyed for providing feedback for each call. Additionally, the deployment of VSD requires the following: 1) it requires changes in the handset and signaling message for getting the response from the callee, and 2) requires a process for learning of responses of a callee towards certain subscribers under different situations, and 3) it also requires interaction among multiple modules which probably would increase the call setup delay.

Caller-REP is different from both Call-Rank and VSD. However it shares some similarity with the Call-Rank that is automatic computation of direct trust between subscribers from the CDRs. Caller-REP collectively utilizes call duration, call rate and out-degree for computing direct trust between subscribers. Additionally, it considers call and social features in both directions incoming and outgoing. The direct trust computed in this way would result in a small direct score for the subscriber having huge number of unique callees with a large number of short duration and non-repetitive calls. The reputation of

subscriber is computed from power iterative method with different initialization procedure. The subscriber having high out-degree would result in a small initial reputation scores and vice versa. In Caller-REP, spammers will get a small reputation score even if they have made and received some good duration calls from their callees. This is because of non-repetitive calling and huge number of recipients of spammers. In perspective of bypassing Caller-REP, spammers need to control call duration, out-degree and call-rate to by-pass the system which is a costly process. In perspective of deployment, Caller-REP does not require any change in the infrastructure for authorization and signaling messages. Additionally, it does not require interaction with the subscribers at any stage of its decision process. However, it can be easily implemented with CAPTCHA or Turing test system as a solution for introducing new subscribers and a way of minimizing the false positive rate.

## 4.4 Experimental Methodology

In this section, we provide an overview of method used for generating the synthetic data-set and the evaluation criteria used for the evaluation of Caller-REP system.

### 4.4.1 Synthetic Data-Set

No VoIP or telecommunications CDRs data-set exists for the testing of proposed detection system. So to test our proposed SPIT detection system, we use synthetic data-set generated by randomly simulating the social behavior of spammers and non-spammers. We considered well-known graph models from the literature. In real telecommunication networks, users exhibits power law degree distribution with  $\alpha$  between 2 and 3 [NEW05a], [NGD+06]. Specifically, the in-degree is between 1.5 and 2, and the out-degree in between 2 and 3. We generate the data set using following: 1) a power-law graph model is generated for subscriber's in-degree and out-degree distribution, 2) a Poisson distribution is used for modeling the call-rate between users in a graph network; and 3) an exponential distribution is used for the call duration between users. We generated the data-set for one service provider regardless the number of proxy servers within the service provider's network. We repeated simulations for at least ten times for different network sizes and percentages of the spammers. The data is generated for 10 days. In evaluation, we provided the average values for specific evaluation metric. We generated data-set for two types of subscribers: spammers and non-spammers. The detail of the data-set is as under.

We have used a power law distribution for the out-degree distribution of the legitimate users with the  $\alpha$  as ( $2 < \alpha < 3$ ) shown in equation 4.4. The call graph is then generated from the power law out-degree distribution using mechanism provided in [GKT+10].

$$p(\text{OutPartners}_S = x) = kx^{-\alpha} \quad (4.4)$$

The legitimate subscribers interact with two types of callees: strongly connected callees and weakly connected callees. We divide the callees of the subscriber according to these groups with few callees in a strongly connected group and others in weakly connected groups. The legitimate subscribers usually have a high calling rate within their strong social group, and moderate calling rate to callee outside their social group. In our simulation settings, the legitimate subscriber follows the Poisson distribution [CHA2015] with the call rate of mean  $\mu = 3$  calls (equation 4.5) with the strongly connected group and  $\mu = 1$  with weakly connected group. Additionally, the legitimate subscriber also receives calls from his strongly and weakly connected groups with the same out-going call rate to respective group.

$$\text{CallRate}_{SR} = \frac{e^\mu \mu^x}{x!} \quad (4.5)$$

Similarly, legitimate subscribers have long duration calls with callees within their strong social group, and average or short duration calls with callees outside their social groups. In our simulation setting, call duration exhibits an exponential distribution [YL07], [JMN+13] with average holding time  $\mu = 360$  seconds (equation 4.6) for strongly connected group and  $\mu = 120$  seconds for weakly connected group. The subscriber also received calls with the same call duration distribution from weakly and strongly connected groups. The out-going and incoming call rate of all users is shown in Figure 4.3.

$$p(\text{CallDuration}_{SR} = x) = \mu e^{-\mu x} \quad (4.6)$$

SPIT callers exhibit different calling behavior than the legitimate subscriber. They usually have large number of recipients in weakly connected group with only few callees in their strongly connected groups. A SPIT caller tries to reach huge number of callees, receives a small number of incoming calls and has short duration for his called and received calls. As such, SPIT caller follows different distributions from the legitimate callers. In simulations, the out-degree of each SPIT caller is uniformly distributed between 20% and 60% of the total number of users in a network. The average call rate of a SPIT caller with the strongly connected group is between 2 and 3, whereas it calls weakly connected recipients only once. The incoming call-rate for SPIT caller is very low and the probability

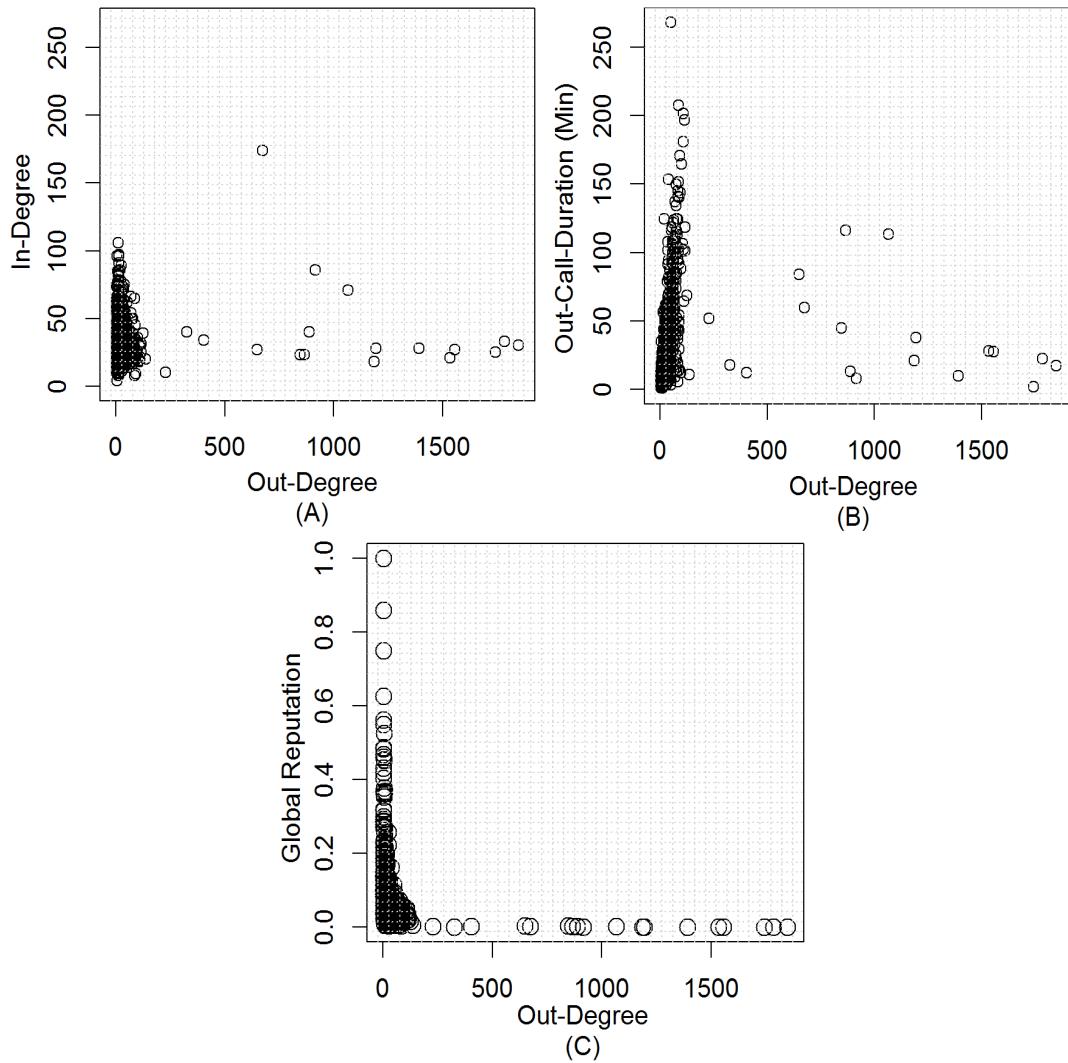


Figure 4.3: Caller Distribution: A) Caller Out-Degree to In-Degree; B) Caller Out-Degree to Out-Duration; C) Caller Reputation to Out-Degree Using Caller-REP.

of receiving a call from recipient of any group is set to .2. The exponential call duration is used for the spammers out-going call duration for both recipients groups with hold time of 180 seconds.

Call statistics for one of our experiment is shown in Figure 4.3. The comparison of caller's out-degree with the in-degree is shown in a Figure 4.3.A, while Figure 4.3.B shows the average call duration vs. the caller's out-degree. Caller-REP assigns small reputation values to callers with high out-degree, high number of short duration calls, and non-repetitive calling behavior; this can be seen in Figure 4.3.C for out-degree.

Prediction/Actual	Spam	Not-Spam
Spam	True Positive (TP)	False Positive (FP)
Not-Spam	False Negative (FN)	True Negative (TN)

Table 4.1: Confusion Matrix.

#### 4.4.2 Evaluation Metrics

We evaluate the performance of Caller-REP against the metrics commonly used in information retrieval and machine learning. The metrics include the followings: true positive rate, false positive rate, and accuracy, and are computed from the confusion matrix presented in Table 4.1.

**True Positive (TP):** True positive rate measures the proportion of spammers that are correctly identified as spammers.

$$TPRate = \frac{TP}{TP + FN} \quad (4.7)$$

**False Positive (FP):** The false positive rate is the proportion of legitimate subscribers incorrectly classified as spammer.

$$FPRate = \frac{FP}{TN + FP} \quad (4.8)$$

**Accuracy (ACC):** The accuracy is the proportion of true results (both spammer identified as spammers non spammer identified as non-spammers) among the total number of spammers and non-spammers.

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \quad (4.9)$$

A well designed SPIT detection system requires to achieve high detection accuracy i.e. correctly classifying spammer as spammer and non-spammer as non-spammer. Blocking a legitimate subscriber would not only annoy legitimate subscribers but also affect the revenue of service provider. On the other hand, allowing SPIT callers to call would not only annoy callees but also affect the productivity of the network.

### 4.5 Performance Evaluation

In this section, we evaluate Caller-REP for the evaluation metrics presented in section 4.4.2 and compare its performance with the performance of Call-Rank. In particular, we

evaluate how fast the system is able to distinguish spammers and non-spammers. In Figures (5.5, 5.6, and 4.6), label "YYY-ZZ%" represent the number of legitimate subscribers (YYY) and the percentage of SPIT subscribers (XX) in a simulated network.

### 4.5.1 True Positive Rate

The first experiment examines the effects of percentage of spammers and non-spammer on the true positive rate of the Caller-REP system. The percentage of spammers is varied from 10% to 30% and the number of legitimate subscribers varied from 100 to 1500 subscribers. Particularly, the performances of Caller-REP has been evaluated for two aspects: 1) how true positive rate of system behaves when number of legitimate subscriber increases, and 2) how true positive rate is effected with the increase of percentage of SPIT callers. The results for true positive rate of both evaluation aspects are shown in Figure 5.5 which plots the fraction of spammers blocked with respect to time.

In a first evaluation scenario that is varying the number of legitimate subscribers with a small percentage of spammers. In this scenario, Caller-REP allows few SPIT callers to pass through the system during first two days, but it starts blocking all SPIT callers with a maximum true positive rate after second day. This is because of the fact that some SPIT callers have small out-degree during first two days and managed to have good duration calls with good number of callees. But, as soon as the out-degree of spammers increases and other call features (duration and call-rate) decrease, Caller-REP start identifying spammers. Specifically, Caller-REP achieves true positive rate of less than 90% during first two days in a network with high number of legitimate subscribers. Over the time, Caller-REP achieves acceptable high true positive rate and eventually achieves almost 100% true positive rate after three days regardless of number of spammers and number of legitimate users. This is because during first two days the behavior of legitimate and non-legitimate subscriber might be same but over the time the legitimate caller develop many strong connections with their callees and spammer develops weak connections with the many callees which is enough to differentiate spammer from non-spammer. This start period or learning period is essential as it help legitimate callers in achieving high reputation scores with the time. For all scenarios the increase in number of legitimate callers, Caller-REP is not allowing any SPIT caller through the system after 3 days.

In the second scenario, the behavior of the Caller-REP system is analyzed for different percentage of SPIT callers. The percentage of spammer varied from 10% to 30% while fixing the number of legitimate user between 100 and 1500. It is expected that system would have high true positive rate when percentage of spammer is small. The true positive rate of Caller-REP decreases with the increase in number of spammers from 10% to 30%

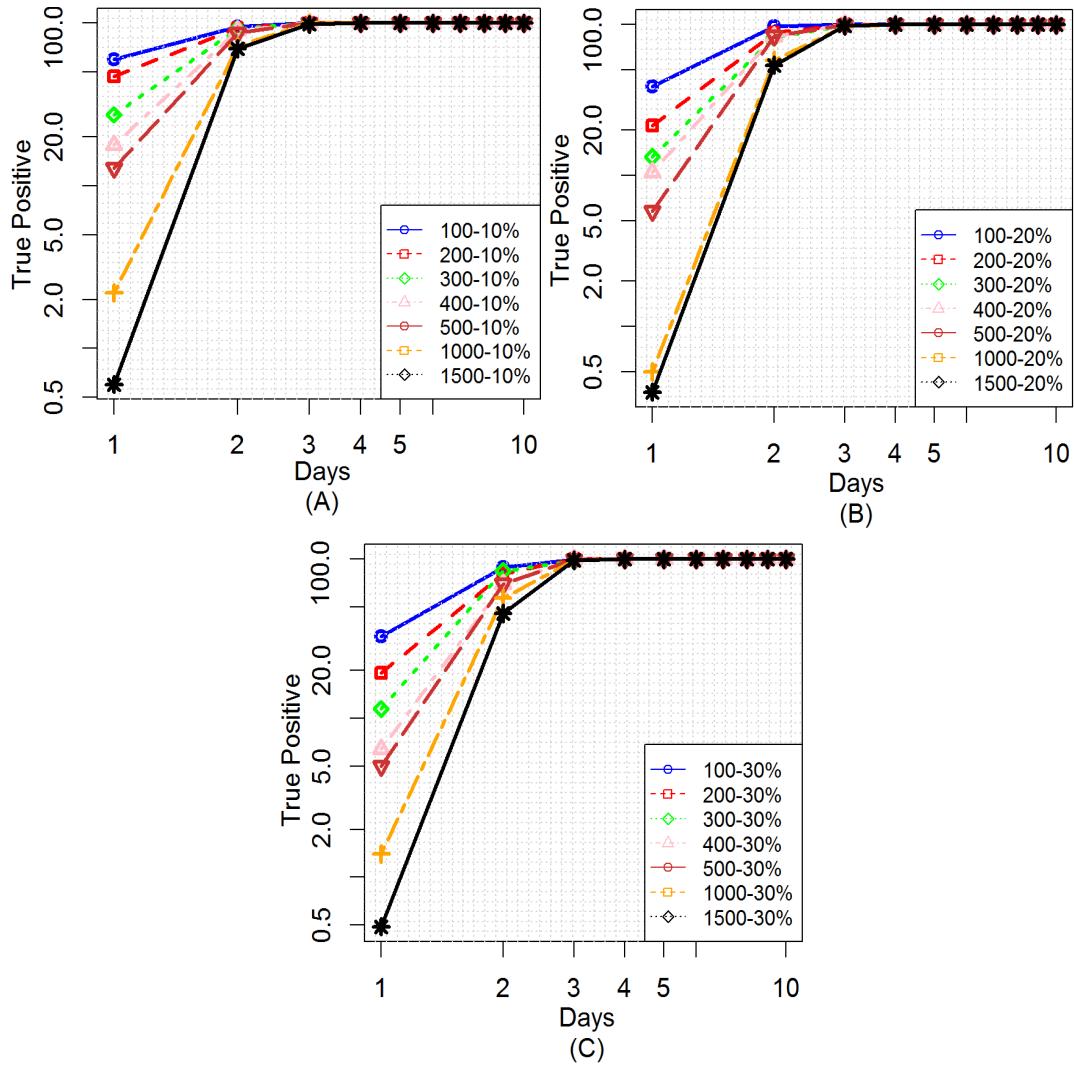


Figure 4.4: True Positive Rate Increases with Time: A) SPIT Rate of 10%; B) SPIT Rate of 20%; C) SPIT Rate of 30%.

as shown in Figure 5.5. Specifically, on a first day, the true positive rate of Caller-REP system decreases by 50% with the increase in number of spammers from 10% to 30%. This behavior is because on a first day, some spammers have small out-degree distribution similar to the legitimate subscriber and develop some relationship with many subscribers. However, over the time as out-degree increases, the true positive rate also increases. The results from Figure 5.5 reveals that Caller-REP is able to achieve true positive rate of more than 90% when the number of spammers is less than 20% and prolong detection to third day when percentage of spammer exceeds 20%. Specifically, Caller-REP is able to block all spammers in three days regardless of number of spammers in the network.

We have also analyzed another critical performance aspect of Caller-REP system that is the time it takes to make the correct classification about the subscriber. The results from

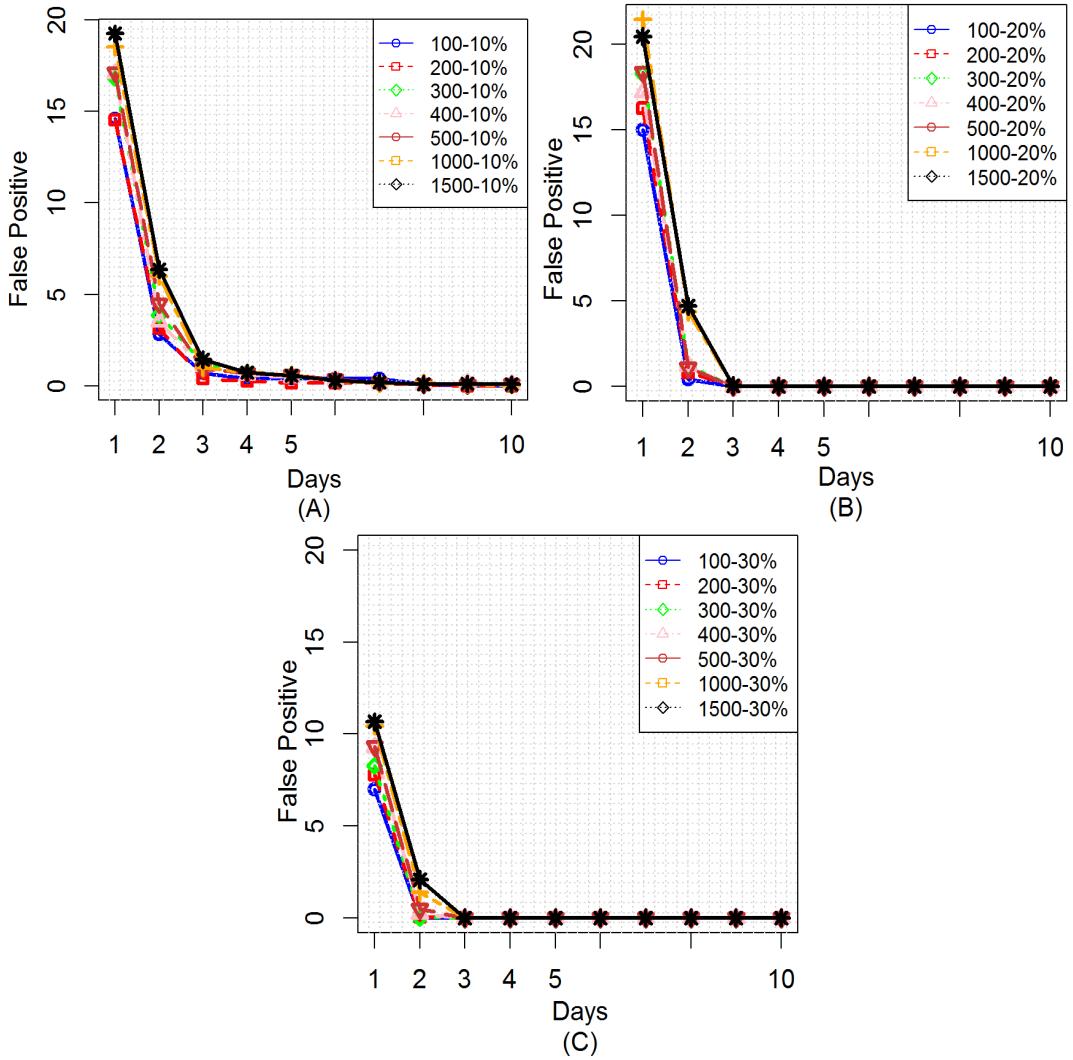


Figure 4.5: False Positive Rate Decreases with Time: A) SPIT Rate of 10%; B) SPIT Rate of 20%; C) SPIT Rate of 30%.

Figure 5.5 show that Caller-REP allows a significant number of SPIT calls to pass through the system on first three days, but it blocks all SPIT caller after third day regardless of percentage of spammers and number of legitimate subscribers. Specifically, Caller REP has high true positive rate for small sized network during first few days rather than large scale network. However, Irrespective of network size and spamming rate, Caller-REP correctly classifies subscribers as a legitimate and a non-legitimate within three days of its initialization.

#### 4.5.2 False Positive Rate

Service provider does not want to block legitimate subscribers for various reasons: first it affects the revenue, secondly it annoys callee expecting calls from some legitimate

subscriber, and thirdly it annoys legitimate subscribers if they bared from calling being honest. A well designed and effective system requires to have small false positive rate without effecting the true positive rate. The false positive rate of Caller-REP for all simulation scenario is shown in a Figure 5.6. The performance of Caller-REP for false positive rate is analyzed for the same performance aspects that we have used while evaluating the true positive rate. In the first simulation scenario where the number of legitimate subscriber increases, Caller REP able to manage a false positive rate of around 15% on first day, but it starts decreasing to less than 2% within three days. This is because some of the legitimate subscribers have short duration calls with many of their callees which results in a small reputation scores similar to that of SPIT callers. However, over the time, if the user behaves legitimately, the reputation scores increases due its strong social connections and if the user behaves non-legitimate, the reputation score decreases over the time due its weak social connection. Caller-REP correctly classifies non-spammer as non-spammer in three days with false positive rate less than 2%.

The analysis for second scenario shows that Caller-REP achieves a better false positive rate in a network with a high number of SPIT callers, as shown in Figures 5.6.B and C as compared to false positive when number of spammer is small. Caller-REP achieves a false positive rate of less than 10% for a SPIT rate of 30% on the first day and eventually decreases to less than 1% on the third day. This means that under high SPIT rate, Caller-REP would not cause revenue loss to the service provider. Additionally, Caller-REP achieves false positive rate of less than 1% within three days for any type of SPIT rate. The false positive rate can be further decreases with the use of CAPTCHA or Turing test to be generated for the subscriber classified as spammers. Additionally, Caller-REP can also be used in combination with social network features like out-degree to decrease false positive in non-intrusive way.

The high false positive rate under small spamming rate and small false positive rate under high spamming rate on a first day is also attributed to the threshold computation. As discussed, Caller-REP is using 25th percentile based threshold for classification which results in a high false positive when spamming rate is small. The threshold can also be adjusted according to SPIT detection policies (require true positive and false positive rates) defined by service providers. In this scenario the 25th percentile threshold is multiplied with a some constant value between 0 and 1. A high threshold would block some legitimate subscribers whereas as small threshold would have small false positive rate. The threshold value need to be chosen in such a way it does not affects true positive rate by high margins.

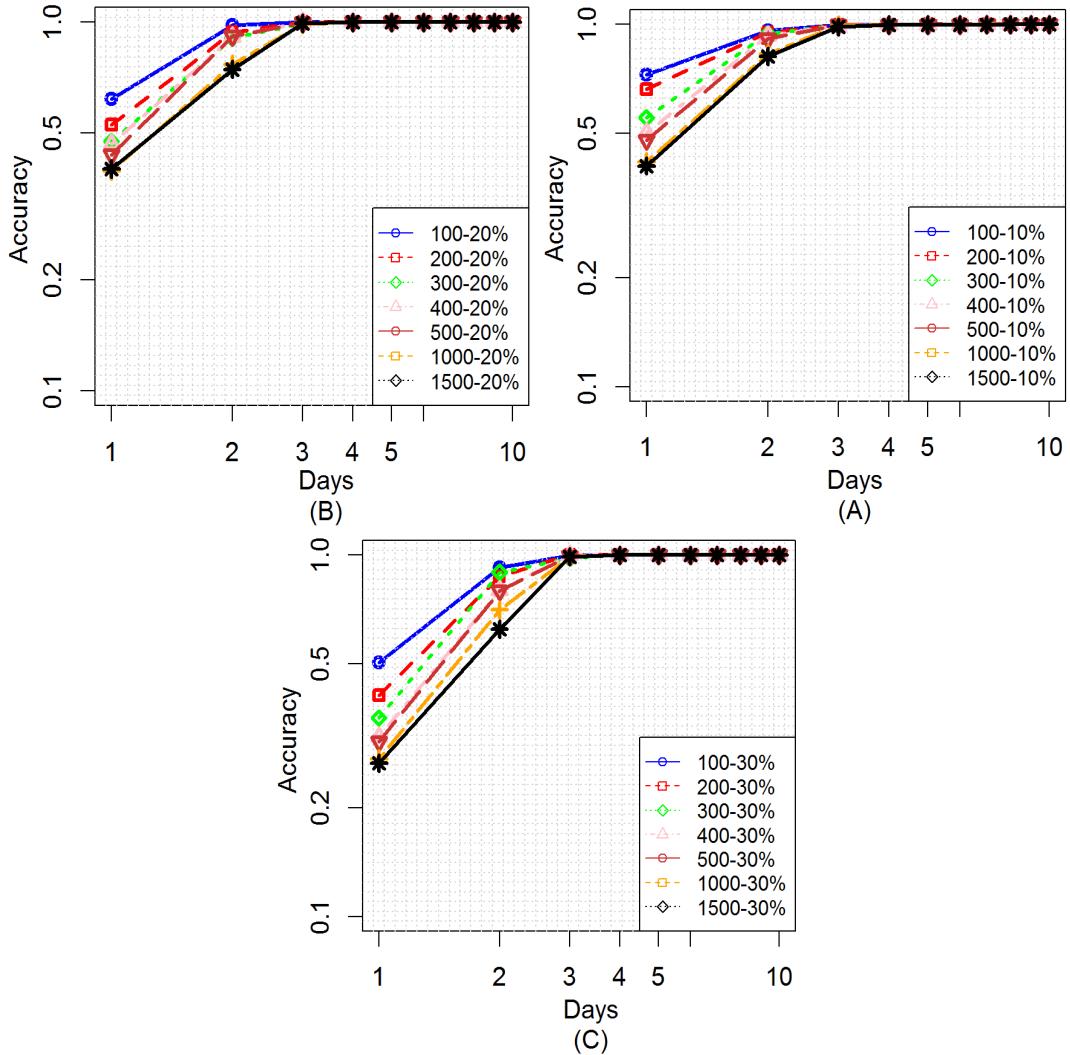


Figure 4.6: Caller-REP Accuracy: A) SPIT Rate of 10%; B) SPIT Rate of 20%; C) SPIT Rate of 30%.

### 4.5.3 Detection Accuracy

The service provider wish to have such SPIT detection system that achieves high false negative (non-SPIT classified as non-SPIT (1-FP)) and high true positive rate. The accuracy metric best characterize the behavior of detection system as it incorporates all four terms of confusion matrix while computing performance of detection system. A small true positive rate results in allowing spammers to reach the subscriber and the high false mistakenly blocks many non-spammers from the subscriber. The True positive rate (TPR) and false positive rate (FPR) of Caller-REP is shown in Figures 5.5 and 5.6 that shows that Caller-REP stabilizes and achieves maximum TP and TN rate within three days. However, TPR and TNR needs to be analyzed together that is best characterize by computing accuracy of the system i.e. systems capability of correctly making decision about the

subscriber behavior.

In first two days, Caller-REP allows many SPIT callers to pass through undetected and blocks some of the legitimate subscribers which results in a accuracy less than 90% when spamming rate is less than 20% and accuracy less than 80% when spamming rate exceeds 20% as shown in Figure 4.6 . Specifically on a first day the accuracy for all spamming rate and number of legitimate user is less than 60% which further improves to almost 100% accuracy in 3 days. Results from Figure 5.5 and 5.6 reveal that the TP rate and TN rate increases over the time which is a positive sign of Caller-REP for not blocking any legitimate subscriber and not allowing any spam caller after 3 days.

The small accuracy during first few days is mainly because of the high false positive rate or small true negative rate. In order to improve the accuracy we need to improve the false positives without affecting the true positive. In view of the above results, we believe that in addition to using a 25th percentile threshold, using some social network features or service provider defined threshold in decision process would be effective in improving the accuracy of the Caller-REP system. The Caller-REP can also relay the call to voice mail box and implicitly analyze the behavior of callee towards the message from voice mail box. It might be possible that subscriber develop only few relations during its introduction periods which results in a subscriber's small reputation score and a sign to consider as spammer. Instead of directly blocking the subscriber having small reputation scores, Caller-REP can collectively use reputation score, threshold and out-degree of the subscriber. If out-degree is extremely small (less than 5) then it would not be sign that subscriber is spammer.

#### **4.5.4 Sparse Subscriber's Network**

In the context of social network, a sparse network is network where the nodes have edges with only fewer nodes from all nodes in a network. In telephone, user develops a scale free power law degree distribution which means that subscriber normally interacts with a small group of callees. This interaction behavior results in a sparse network matrix or sparse network for users. In this section, we analyze the performance of Caller-REP system in a sparse network. We expect, that sparseness would not result in a small reputation scores to subscribers having sparsity because of use of collective use of three call features: call duration, call rate and out-degree. In Caller-REP network sparseness is not the only factor that has an impact on a global reputation and direct trust. A sparse caller only gets bad reputation when along with sparseness it also has low call duration and low call rate.

The evaluation of a sparse network requires a different simulation setup than the simulation setup that has been used for analyzing the TPR and FPR. For analyzing the effects

of sparseness, we created a full network for 11000 users using the method from the section 5.5.1 and then added a percentage of users with a sparse network. We performed simulations for two scenarios. In a first scenario, 25% of callers has less than 10 friends. In this scenario, Caller-REP misclassifies only 11% of legitimate subscribers as spammers. In a second scenario, 45% of subscribers has less than 20 friends and Caller-REP misclassifies only 7% of legitimate subscribers. The results also show that decreasing the degree of sparseness would decrease the FPR. Additionally, the sparseness has no effect on the TPR. On further investigation, the sparse legitimate subscribers that were misclassified as spammer were found to have small average call duration and calling rates.

### 4.5.5 Subscriber Reputation

The reputation of a legitimate subscriber increases and reputation of non-legitimate subscriber decreases over the time. This is because of the fact that the legitimate subscriber has repetitive long duration calls to a large number of their called callees and has short duration calls with only few callees. On the other hand non-legitimate subscribers target large number of callees and managed strong connection with only few callees. This behavior would result in a high reputation score to legitimate subscriber which further increases over the time and small reputation score to the spammers which further decreases over the time. The global reputation scores of spammer fluctuate around his initial reputation score whereas non-spammer show increase in reputation scores from their initial reputation scores. The small reputation score to spammers and high reputation score to non-spammer is because of the collective use call-features (duration and call rate) and the out-degree distribution. The high out-degree with small duration call will largely effect the reputation of the subscriber.

In Caller-REP, it is difficult for spammer to obtain and maintain high reputation scores. This is because the reputation is based on the call interaction, call duration, and the out-degree of the subscriber. The spammer needs not to have good duration bi-directional repetitive calls but also need to control his out-degree in order to have high reputation score but this is not practical in real scenario and spammer would not have benefit from it. A legitimate subscriber, on the other hand, will have a high reputation value due to high number of repetitive long duration calls in both direction and relative small out-degree. Figure 4.7.A shows the reputation score distribution of spammers and non-spammer over the period of 10 days. From Figure 4.7.A, it is clear that reputation of suspect spammer is not increasing much and does not vary over the time, whereas the legitimate subscribers show continuous increase in their reputation scores. The reputation scores for Call-Rank

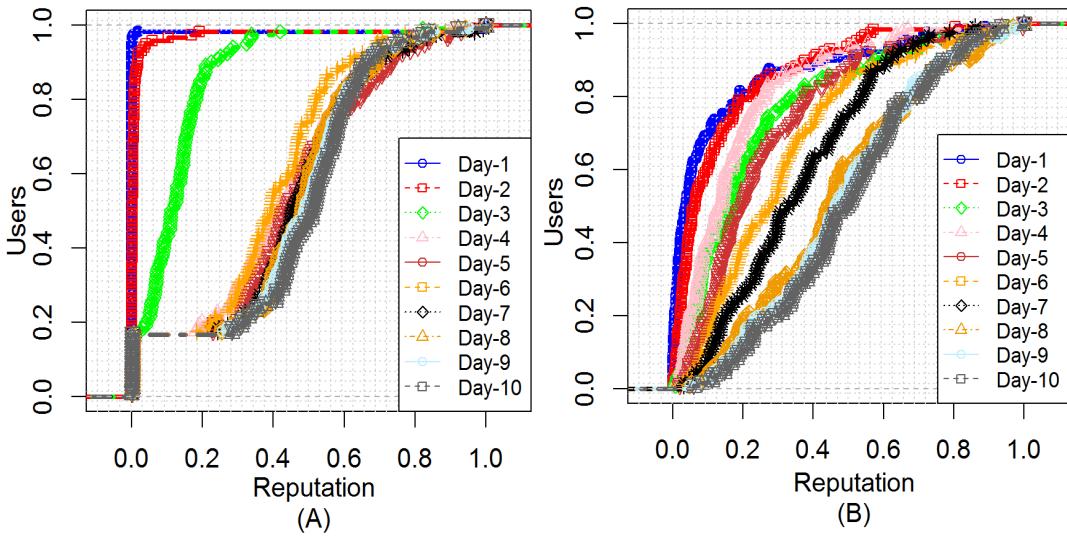


Figure 4.7: Caller Reputation With The Time: A) Caller-REP; B) Call-Rank.

system for ten days is shown in Figure 4.7.B which shows that global reputation of spammers is also increasing if subscriber has high out-degree and some moderate duration calls. From a Figure 4.7.A, we also conclude that 25th percentile could provide better detection accuracy for the Caller-REP than the Call-Rank system.

#### 4.5.6 Caller-REP Vs. Call-Rank

In section 5.4.1 we have presented how Caller-REP is different from other SPIT detection systems. In this section we compare the performance of Caller-REP with its closely related counterpart approach the Call-Rank. The social network of legitimate subscriber becomes strengthen with the passage of time as compared to non-legitimate subscriber having weak social network over the passage of time. The reputations of legitimate and non-legitimate subscribers using Caller-REP and Call-Rank approach is presented in Figure 4.7. The reputation of legitimate and non-legitimate subscribers in Call-Rank increases with the time as shown in Figure 4.7.B. This means that if Call-Rank misses a certain non-legitimate subscriber on the first day, it would not be able to detect this suspected subscriber on next days because of his improved global reputation value. On the other hand Caller-REP would not increase the reputation of suspected non-legitimate subscribers over the period of time instead it decreases over the time. This means that if Caller-REP misses a certain non-legitimate subscribers on the first day then it would eventually detect it on the next days because of its slight change in a reputation values. In terms of detection performance, the Call-Rank achieves a maximum true positive much later then that of Caller-REP system. However, Caller-REP achieves a true positive rate of

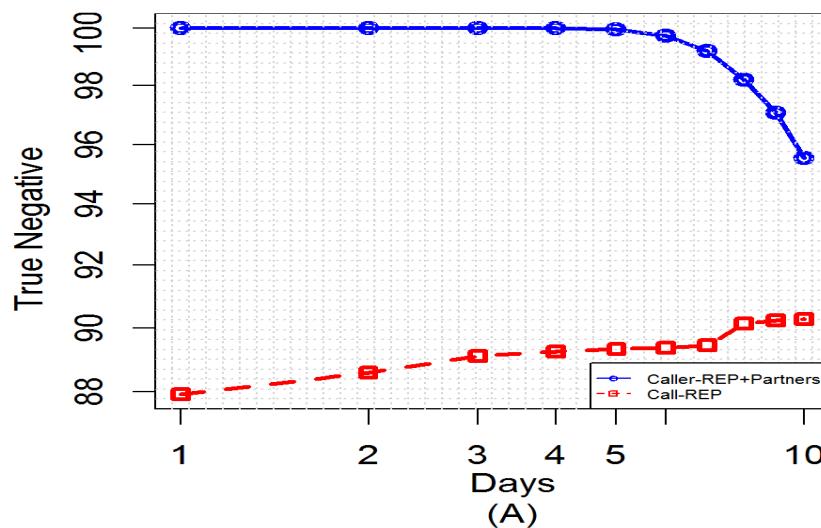


Figure 4.8: Caller-REP Performance Under Legitimate Network.

almost 100% on the third day under any type of spamming network. In terms of false positive rate, caller-REP blocks 10% of the legitimate subscribers, which is much higher than the false positive rate of Call-Rank. However, with time this false positive rate decreases to less than 1% and behaves similarly to a the Call-Rank.

#### 4.5.7 Caller-REP under Legitimate Network

The deployment of Caller-REP in a real VoIP network may face another major challenge when all subscribers in a network are legitimate. There are no SPIT subscribers in this condition. In its original form, Caller-REP may wrongly classify weakly connected legitimate subscribers as SPITter. In order to minimize this misclassification, Caller-REP can be improved using additional social network features and improved automatic threshold detection. The following social network features could be used along with Caller-REP: clustering coefficient, ratio of incoming to total calls, path distance measure and out-degree distribution. The computation of few these features becomes difficult when the subscribers are on different networks and service providers are not willing to share callee internal network structure. However, out-degree and in-degree distribution of subscriber can be easily computed without extra effort and provide information along with reputation score.

Usually the SPIT caller tries to reach a large number of callees and consequently has a more unbalanced out-degree distribution than the legitimate subscribers. In a legitimate network, Caller-REP misclassifies callers with short duration calls and small number of unique callees. The small number of unique callees in a day cannot be taken by itself

as the sign of SPIT callers [SWN12]. Figure 4.8 presents the performance of Caller-REP and Extended-Caller-REP under a legitimate network. In extended Caller-REP we consider callers as legitimate if their number of out-partners is less than 5 even if the subscriber is classified as non-legitimate by Caller-REP. Caller-REP achieves 90% true negative rate (non-SPIT detected as non-SPIT) in a legitimate network. The extended Caller-REP behaves well at start, but its true negative rate decreases to 96% with the time. The decrease in true negative in extended Caller-REP is due to the fact that some legitimate subscribers have low duration calls with few callees and also do not receive calls from called callees.

#### **4.5.8 Caller-REP under High SPIT Rate**

Currently the email spam traffic dominates the total legitimate email traffic. However, in Telephony not many SPIT events have been reported. In future, advertisers will likely start using VoIP as a mechanism for advertising their products. In a high SPIT attack, where the network comprises more SPIT traffic than non-SPIT traffic, the Caller-REP only blocks a limited number of SPIT callers having extremely low reputation values. The positive aspect of Caller-REP under a high SPIT rate scenario and  $\beta$  value of 2 is that it would not block any legitimate subscribers. However, in order to block all SPIT callers, the Caller-REP can be improved by carefully setting the  $\beta$  parameter. We performed experiments for 2000 users, for a varying number of spammers from 45% to 80%, and a fixed  $\beta$  value of 2. Figure 4.9 presents the true positive rate of Caller-REP under high number of SPIT callers and  $\beta$  value of 2. The true positive rate decreases with the increase in a percentage of SPIT callers; however, Caller-REP only allows less than 8% of SPIT callers to make calls under high number of SPIT callers. The Caller-REP allows all legitimate subscribers with a zero false positives. The true positive rate of Caller-REP under heavy SPIT attack also stabilizes with the time and achieves the true positive rate of 96% within 3 days and around 98% in 8 days.

## **4.6 Discussion on Caller-REP**

In this section, we discuss various aspects of Caller-REP system. We first discuss some important characteristics of Caller-REP system and discuss how it can be bypassed by the spammers. We then discuss deployment issues of Caller-REP in a real VoIP network.

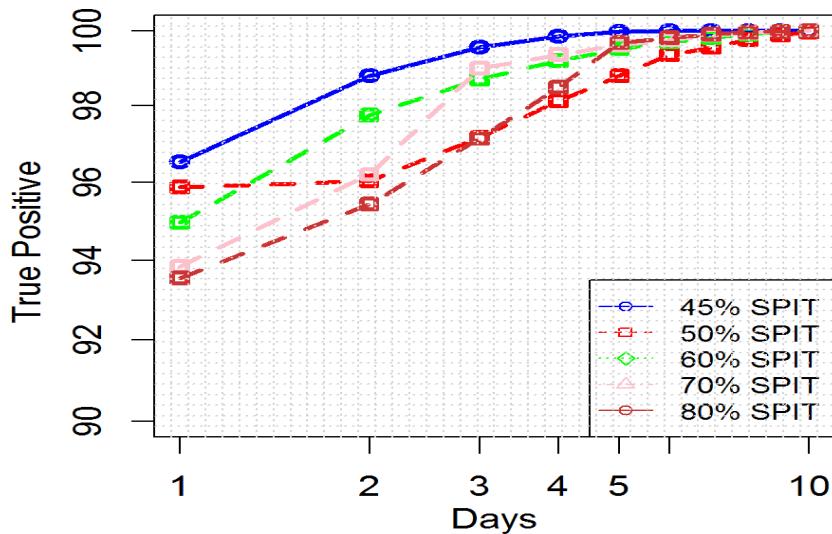


Figure 4.9: True Positive Rate Under High SPIT Traffic and  $\beta=2$ .

#### 4.6.1 Features of Caller-REP

Caller-REP can be by-passed by the SPIT caller under certain conditions. The SPIT caller is able to reach the callee only if he managed to have good trust with a large number of called callees. This can be achieved by having long duration and repetitive calls to a large number of callees. Caller-REP use three features for a direct trust computation and out-partner is the only one under the subscriber's control. The other features that are call duration in both direction and repetitive behavior depends on callee behavior towards the subscriber. The subscriber can be passed through the system even it has short or no duration incoming calls provided that he has controlled out-degree with repetitive and long duration calls with his called callees. There are some service providers offering flat-rate packages for the domestic and international calls and is not profitable for the service provider when abused by telemarketers or spammers for large number of calls. These flat-rate telemarketers need to be distinguished from the normal flat-rate subscribers for profit. The legitimate flat subscriber usually tries large number of callee in a short which might have out-degree distribution same to low rate SPIT caller, but the legitimate flat subscriber usually has social circle quite different from the SPIT caller. The social circle of non-SPIT caller would result in a repetitive calling behavior and good call duration to a large number called callees. This calling behavior of non-SPIT flat caller distinguish from the SPIT caller even the SPIT callers are not receiving any call from their called callees. However, Caller-REP would block flat-rate legitimate subscriber having large number of called callees and small call duration.

Caller-REP blocks such legitimate subscribers having high out-degree and short duration non-repetitive calls to his called callees. As an example, an employer wants to convey

short duration message to all his employees. This calling pattern would result in a high out-degree and small duration calls to their called callees. Similarly, the organizations offering services for the betterment of society also have high number of unique callees with short or large duration calls and are considered as SPIT caller by Caller-REP. However, these callers are very few in number and can be assigned high trust scores to be passed through the system. There are other situations where Caller-REP also block legitimate subscriber: for example when a job seeker tries to call many companies for job with a less or few incoming short duration calls and. This can be overcome through personalized Caller-REP.

It is also possible that a legitimate subscriber may has short duration calls to a large number of his called callees and may be considered as SPIT caller because of call duration similarity with the non-legitimate subscribers. However, the out-degree, repetitive calling behavior, number of incoming calls to such caller and duration of these incoming calls to the caller are different from the SPIT caller. No single feature such as short call duration or small out-degree proves anything specific or absolute about the nature of a subscriber. The call duration, repetitive calling behavior and number of unique callees of a subscriber are the three features that have been used collectively in a Caller-REP and affect the reputation of subscriber across the network. The subscriber having short duration calls would have high reputation if he calls his callee frequently and receive calls from them no matter how many unique callee he has. The Caller-REP also learns automated threshold and would only block such subscriber having short duration calls with large number of unique callees, has non-repetitive calling behavior and small number of non-repetitive incoming calls.

#### **4.6.2 Sybil Attack**

In a Sybil attack subscribers achieve high reputation scores by acquiring large number of identities, using them to gain high reputation in order to make negative impact on other users with spamming. The detection of Sybil network depends on the cost imposed while acquiring a new identity. The Sybil subscribers gain maximum advantage by creating large number of fake identities, if cost on creating a Sybil subscriber is small. The strong authentication and high cost for creating the new identity could limit the Sybil attack. In VoIP or telephony, acquiring a new identity is not only easy but cheap. A Sybil user can easily make calling network between all his identities by making calling network between his identities before finally making calls to other users. Caller-REP can impose two methods for Sybil subscriber: firstly, it imposes some cost for new identity but this approach would also limit number of legitimate subscribers. Secondly, it adopts social patterns of

Sybil subscribers. Caller-REP is based on the assumption that Sybil subscriber would have limited number of strong relationship to real users. The Sybil spammer gets some profit when it calls to a large number callee outside its Sybil network, which normally results in small reputation scores for the Sybil subscriber and a possible sign of spammer. For the social pattern based Sybil detection, Caller-REP adopts mechanism of not using trust values from the callees which are identified as spammers or has extremely small trust values. The effect of Sybil subscriber can also be minimized by considering the combination IP-address and user identities during the creation of subscriber's trust network but it is possible also limit legitimate caller for example a legitimate call center or large organizations.

### 4.6.3 Betrayal Subscribers

Betrayal subscribers are those callers who start spamming after creating trust link with the large number of subscribers. The large number of trusted links would result in a high reputation score for such callers. This scenario can happen in two ways: 1) Betrayal subscribers deliberately achieve some good reputation scores by creating trust link with other spammers or legitimate callers, and 2) the spammer compromised or spoofed identity of the legitimate subscribers. The performances of the detection systems normally degraded under these calling behaviors and allow Betrayal spammer for relatively long time periods. It is utmost important that detection system must be adoptable to the changing calling behavior of Betrayal callers. Though the Betrayal callers can have high trust and reputation score for some time period, but it would not be possible for them to maintain it when they target large number of callees. A change in the calling behavior and large number of small duration calls to a large number of callees would drastically ruin their achieved high reputation scores.

Caller-REP blocks Betrayal callers as soon as these callers start spamming the subscribers of the network. Specifically, Caller-REP utilizes subscriber's call patterns in a specific time window for computing his direct trust and reputation. When a reputed subscriber starts acting as a SPIT caller, his direct trust values start decreasing due to his abnormal call patterns to a large number of unique callees. Caller-REP makes it difficult for the suddenly converted untrusted subscribers to maintain their high reputation for the long time periods because of collective use of number of social and call features for computation of global reputation.

#### **4.6.4 Addition of New Subscriber**

When a new subscriber wants to join the service provider network, the existing network subscribers do not have any information about the joining and reputation of the new subscribers. Since new subscribers are new to the system they must be introduced to the network. The new subscribers can be introduced in two ways: 1) by considering all new comers as trusted and 2) a new user is not trusted. The first approach allows spammers to pass the system and second approach might block some legitimate subscribers before building their reputation. The first approach with some additional constraint is suitable for introducing new subscriber to others. Calls to small number of unique callees is not the sign that subscriber is spamming and can be used as threshold for allowing caller for some fixed number of callees. In this way, the subscriber will get a fair chance of improving their reputation score in order to have future calls. The selection of threshold for the fixed callees should neither be small nor be large as a small threshold would block legitimate subscribers and large threshold would allow spammers for considerable several callees.

Caller-REP expects that subscriber should maintain some level of trust with other subscribers in order to have good reputation and future communication. Legitimate subscribers normally have repetitive and long duration calls with their friends and family members over the time which results them to high reputation scores. However, SPIT callers make large number of calls as soon as they join the network which normally results in a high out-degree and small reputation scores. In order to provide fair chance to subscriber for his introduction in the network, Caller-REP allows new subscriber to remain unchecked as soon his number of unique callees is less than 5. As the number of callees exceeds 5, Caller-REP starts computing reputation of the subscriber and classifies him as spammer or non-spammer considering his reputation score. Imposing this form of introduction ensures that the new subscriber gets fair chance of making social link with the callees. Caller-REP can also be implemented with CAPTCHA test as a solution for introducing new subscribers but this is not only intrusive but also requires some system resources from the service provider.

#### **4.6.5 Deploying Caller-REP in a Real VoIP Network**

The VoIP service provider authenticates and authorizes subscriber before allowing them to call the callees. The service provider is responsible for implementing call routing, billing policies, and logging of call transactions. The Call detail records are generated for every successful and failed call at the proxy server and are periodically parsed to the CDR server for billing and backup. In a real VoIP service provider, a Caller-REP can be placed on any of these locations: 1) the VoIP proxy server or call processing engine,

2) the CDR server, and 3) deployed as a standalone network component. In any form of deployment, Caller-REP requires interaction with CDR server and requires adjustment of two parameters: the  $\beta$  parameters and a threshold for the number of unique callees. These two parameters control the true positives and false positives. A small value of  $\beta$  parameter and a high number of out-partners may minimize the false positives but might result in some revenue loss. The number of unique callees can be fixed to 5 as such a low call-rate cannot be the behavior of a SPIT caller. The network service provider can choose small or high threshold for the number of unique callees. The high threshold on the number of unique callees may allow SPIT callers to reach relatively large number of subscribers and small threshold probably block legitimate callers having small duration calls.

The service provider also wants to maximize his profit and may not be willing to block all SPIT callers except the ones having very small reputation scores. Consider a Telemarketer who has fixed calling rate from the service provider but makes unsolicited calls to a large number of network subscribers. Blocking these callers from calling would result in a revenue loss to the service provider. The VoIP service provider can set  $\beta$  parameters according to his requirements and SPIT detection policies. The high threshold would block all suspected SPIT callers and would also results in some false positive which is again a loss of revenue to the service provider. The low threshold would block maximum number of SPIT caller's and allow few SPIT caller to pass through the system which normally results in a no false positives and no loss of revenue. The service provider can also use reputation scores along with the callee consent and contractual agreement. In this case the service provider first learn the callee threshold for his caller's reputation and allow only those caller to call the callee having reputation score greater than callee learned threshold.

#### 4.6.6 Call Setup Delay

The primary use of VoIP telephony is to satisfy customers and provide committed QoS (Quality of Service) comparable to traditional circuit switched telephony. The placement of security system on call processing engine would introduce call setup delay and may not be acceptable for service providers. Call setup delay is the delay between when the caller enters the last dialed digit and when he receives the ringing tone. In a SIP-based VoIP network, call setup time can be taken as the time interval between when the proxy server receives the invite request and when it responses back with the "180 ringing message" to the caller. Call setup delay is the only metric that can be affected by the placement of Caller-REP in a network. Caller-REP is designed to work independently from the call processing engine thus would not add any noticeable delay. The call processing

engine requests Caller-REP for the reputation of a caller at the same time it is processing routes for the call and getting billing policies from the billing server. As such this parallel processing of getting reputation of a caller would not introduce any observable delay to the caller call request.

## 4.7 Conclusions

The subscribers call transactions are recorded in a CDR that can be used to generate a weighted social call graph between subscribers. The collective use of social network and call features for computing reputation of the subscriber can be highly effective for blocking SPIT caller. In this chapter, we presented a Caller-REP system – a SPIT detection system that automatically identifies spammers in a VoIP network without content processing and user involvement. Caller-REP considered variety of social network and call features for computing direct trust between subscribers, and global reputation of subscribers is computed by aggregating direct trust score of subscribers. The approach results in a small reputation scores to the SPIT callers and high reputation scores to the non-SPIT callers. The automated threshold is then computed from the global reputation score below which the subscriber is considered as the spammer. We evaluated the proposed approach through series of experiments consisting of different percentage of SPIT and non-SPIT callers. The results show that the Caller-REP is an effective and efficient system for detecting SPIT caller under any type of network conditions. It achieves true positive rate of almost 98% and false positive rate of less than 1% within three days. The privacy of subscriber is also being full protected as VoIP service provider is neither disclosing sensitive information of caller to reputation engine nor providing any information to the called callees. Finally, we compared the performance of Caller-REP with Call-Rank. In our experiments, Caller-REP outperformed Call-Rank in terms of both detection accuracy and detection time. The Caller-REP is specifically designed to be used within existing VoIP infrastructure without any change in a network architecture and call flow. The system can also be implemented with other existing solutions for improved accuracy and early detection.

Based on these encouraging results, in a next chapter we present collaborative SPIT detection system that involve collaboration among service providers for the early identification of spammers making low rate spam calls to many service providers. In collaboration, the VoIP service providers exchange the reputation scores of a subscriber to the central repository, which in turn decides the nature of a subscriber on a behalf of the

service provider. We believe that this collaboration would further minimize the detection time and achieves high accuracy.

# Chapter 5

## COSDS: Blocking Spammers with Information Sharing across Multiple Service Providers

### 5.1 Introduction

Telecommunications service providers (SPs) can deploy standalone spam detection systems [DK05], [KD07], [WBS+09], [BAP07], [AM12], [AM13] within their network for protecting their subscribers from unsolicited calls and SMS. The standalone detection systems 5.1(a) consider data from one source for analyzing the calling behavior of subscribers within service provider. Spammers can evade these standalone systems by making a large number of spam calls in aggregate to recipients of many service providers without overwhelming any single service provider with the spam calls. By doing so, spammers remain undetected for a longer time period within the service provider, since service provider is not receiving large number of calls from the spammers that flagged them as spammers. An effective solution to detect low rate spammers requires monitoring of behavioral patterns of subscribers across multiple SPs.

Obviously, collaboration and information sharing can be an effective way to block such spammers making low rate spam calls to recipients of many service providers. However, there are two key challenges in collaborative spam detection: firstly, what information should be exchanged during the collaboration process; and secondly, to whom this information should be made available. Collaboration will probably achieve better detection accuracy and time than the standalone detection systems but its performance depends on the amount of information exchanged among collaborators. The collaborative solution can be either the distributed 5.1(b) or the centralized 5.1(c): distributed - where

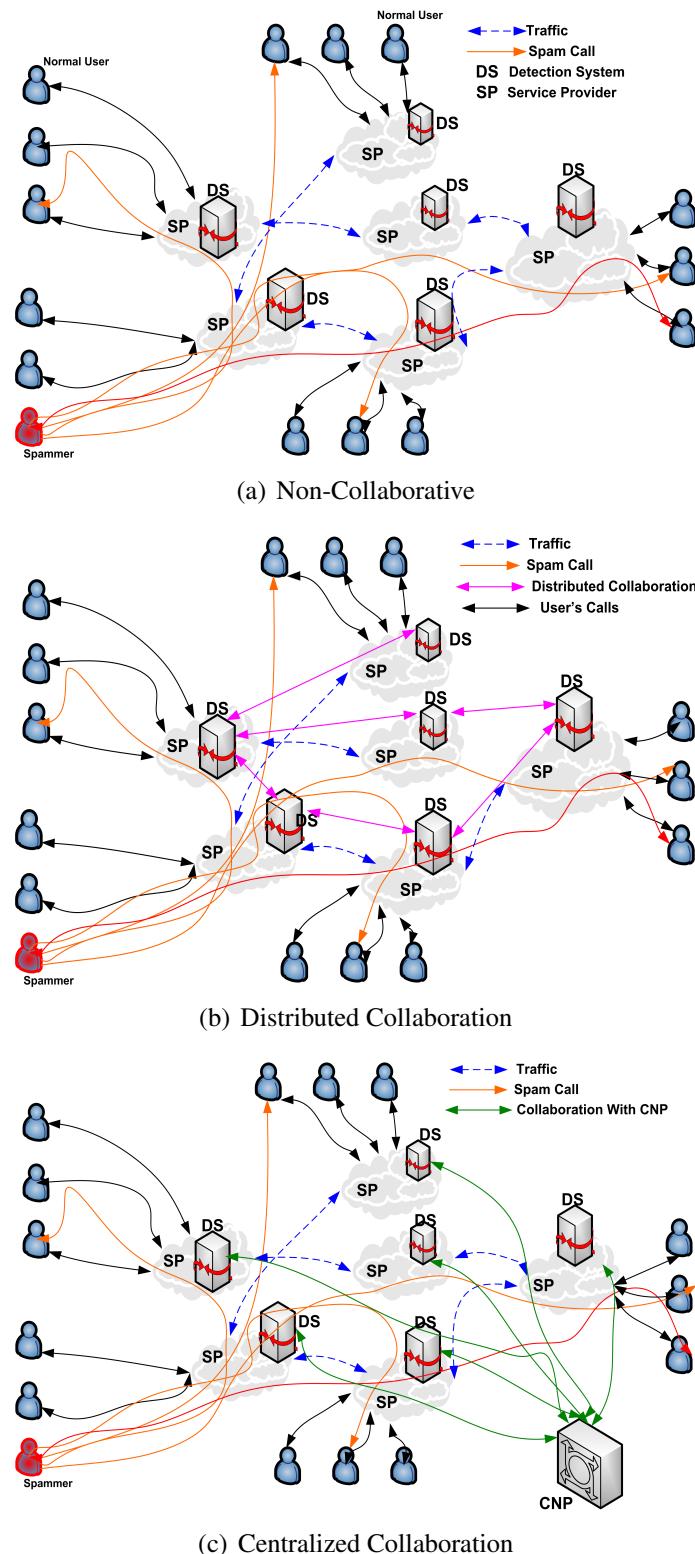


Figure 5.1: Collaboration Methods: A) Non-Collaborative; B) Distributed Collaboration; C) Centralized Collaboration.

the information from each SP is shared and processed in a completely distributed fashion without a central coordinator; centralized - where all information from SP is reported to the single centralized location for analysis. The distributed data aggregation and analysis lack privacy protections. The service providers are not willing to share operational data of their customers with each other, because they are business competitors and are concerned about compromising the privacy of their customers. Generally, better detection accuracy is expected when collaboration is achieved through the exchange of complete call records but at the cost of system and network resources. Moreover, SPs are likely to be reluctant in a direct exchange of CDRs to peer SPs or the trusted Centralized Respository (CR) because CDRs contain subscriber's private information as well as operational details of their network. SPs may be more comfortable in exchanging summarized information to the trusted CR rather than exchanging information directly with other SPs. The exchange of summarized information could deteriorate the system performance in terms of detection accuracy and time but it does not require extensive network resources and exchange of call records. A key challenge in the design of a collaborative SPAM detection system is to achieve high detection accuracy and minimizes the detection time without compromising computation resources and without the exchange of CDRs.

To address the above challenges, we propose a Collaborative Spit Detection System (COSDS) for an accurate and early detection of SPIT subscriber, which is based on collaboration among many autonomous SPs. The major feature of COSDS is that it does not require direct collaboration among SPs. Instead the collaboration is carried out with the exchange of summarized information with the trusted CR thus reducing the network load. In particular, each SP submits the Local Reputation (*LR*) scores (summarized information) of their subscribers to the trusted CR. These reputation scores represent the behavior of subscribers within the SP network and have been computed from subscribers past call transactions within the SP. The CR is responsible for the computation of global reputation of subscribers by aggregating the received *LR* scores and deciding about spamming behavior of subscribers. The CR responds collaborating SPs with the *GR* scores and decisions about subscribers, which also allows each SP to act independently against spammers. In COSDS, only a summarized information i.e. *LR* scores are sent to the trusted CR which are not resource demanding. Each collaborating SP interacts directly with the CR and requires only two transmission cycles for getting *GR* of their subscribers i.e. one cycle for sending *LR* to the CR and one cycle for receiving *GR* from the CR. Additionally, the use of trusted *CR* and exchange of summarized information further likely to convince SP to take part in a collaboration.

We evaluate our system using synthetic data that we have generated through models of spammers and non-spammers social behavior. The evaluation has been performed for dif-

ferent performance metrics and for different percentages of spammers and collaborators. We demonstrate that collaboration among SP outperforms standalone detection systems in terms of detection accuracy and detection time. Specifically, for a network having a large number of spammers, COSDS managed to achieve zero FP rate and blocked all spammers within 3 days. The results also reveal that COSDS achieves detection accuracy that is comparable to that of a system where collaboration is carried out through the exchange of call detail records. COSDS approach is fast, requires small communication overhead and only requires a few iterations for the reputation convergence within the SP.

The proposed approach is an extension of the SP level SPIT detection system presented in Chapter 4. In this chapter, we establish cooperation among SPs and focus on defining the components and mechanism for collaborative SPIT detection. This enables early and accurate detection of the spammer while considering the *LR* scores of the subscribers across many collaborating SPs.

In a summary, the contributions of this chapter are:

- The design of a collaborative SPIT detection system that incorporates collaboration from multiple autonomous SPs for an early identification of spammers distributing low rate spam calls to recipients of many SPs. Each autonomous SP is capable of processing locally recorded call transactions of their subscribers for computing *LR* scores of subscribers, which are then sent to the trusted CR. The trusted CR computes *GR* of subscribers by aggregating the reputation scores and makes meaningful decisions about behavior of a subscriber as a spammer or a non-spammer. The exchange of summarized local reputation scores not only convince SP to be a part of collaboration process but is also not resource demanding regarding network and system resources. The proposed centralized design and exchange of summarized information further ensures the privacy protection of subscribers within SP as well as at the CR.
- A detailed evaluation has been performed on the synthetic CDRs. Particularly, we evaluated the system for different number of collaborators, different percentage of spammers and for the following metrics: true positive rate (TPR), false positive rate (FPR) and accuracy. We also compare the performance of COSDS to a system where collaboration is carried out through the exchange of CDRs or direct trust scores with the CR. In addition, we also evaluate subscriber's privacy aspects within the collaborating SP and at the centralized repository for different auxiliary information known to adversary.

The chapter is structured as follows. In Section 5.4, we describe architecture of collaborative SPIT detection system and design options for the collaboration. Additionally,

Section 5.4 also provides discussion on deployment challenges of COSDS in the SP network. The experimental setup is presented in Section 6.4 and detail evaluation for different performance metrics is presented in Section 5.6. In section 5.7 we discuss features of collaborative system and then conclude the chapter in Section 5.8.

## 5.2 Limitations of Stand-alone Detection Systems

Stand-alone SPIT detection systems are currently major systems for thwarting SPIT subscribers. These systems are typically placed within one SP and consider only locally recorded data within the SP for deciding about behavior of the subscriber as a spammer and a non-spammer. Since there is no cooperation among SPs, no data from SP is passed to other SPs except call handling messages. Standalone anti-SPIT systems may have high false negative rate and prolonged detection when spammers are making low rate spam calls to recipients of several SPs without making large number of spam calls to any single SP. In particular, stand-alone systems could manage to detect low rate spammer over time (after receiving enough number of calls) and when the number of spam calls from the same subscriber spikes. However, this detection is too late as spammer has already reached to a large number of subscribers in a particular SP and across several SPs. The prolonged detection is because of unavailability of information for making reasonable decision about the sender. The stand-alone systems can improve their detection capability by combining several stand-alone detection approaches into a single multistage system or asking subscriber for solving the CAPTCHA test. However, these implementations have following limitations. First, CAPTCHA involves subscriber for solving the challenge that is not only resource intensive but is also intrusive to the subscribers. Second, multistage systems require call request to pass through many detection components thus would increase the call setup delays. Third, multistage systems still require a relatively large number of calls from the same subscriber for making the final decision about the subscriber and still allows spammers to reach several subscribers.

## 5.3 Motivation

Existing SPIT detection systems classify subscriber as spammer and non-spammers based on the call patterns of subscriber observed at a single service provider. A low rate spam subscriber that distributes spam calls across many SPs may evade the stand-alone detection systems. However, for financial benefits, a low rate spammer makes a low rate unsolicited calls to recipients of many SPs and his calling behavior remain same across

all target SPs. Thus, observing calling behavior of user across multiple SPs could help in early identification of spammers that are responsible for making large number of spam calls. The existing collaborative anti-SPIT systems [SS09] though involve collaboration among SPs but it only rates detection capability of spam detection system placed in home network of the subscriber. The COSDS approach is different from [SS09] in a sense that it computes reputation of end users instead of computing reputation of detection system placed in a home service provider of the end user. Moreover [SS09] requires changes in the call setup messages to incorporate tags that are exchanged between collaborators, whereas COSDS does not require any change in the network architecture and call setup messages.

To increase the detection accuracy and reduce detection time, it is utmost important to establish a collaboration among SPs for computing aggregate reputation of end-users. The effectiveness of collaborative anti-SPIT system mainly depends on the amount of information being exchanged in a collaboration process and has challenges of privacy protection, communication overheads and system resources. There is a strong need to have a collaborative system that fulfills following conditions. 1) Collaboration among SPs needs to be carried out without establishing a direct trust relationship between collaborators. 2) The information used for the collaboration should not be resource intensive regarding network and system resources. 3) The information exchanged should not contain any sensitive information that could be used by the adversary to infer the relationship network of users. 4) The design system should have high true positive rate and small false positive rate.

## 5.4 Collaborative SPIT detection System

We consider four assumptions: 1) people calling behavior can change over time (they add or remove links, have different call behavior with family and friends etc.) [SMS+08]; 2) the calling behavior of legitimate subscriber is different from that of spammer [BSG+11], [CMP+13], [DTN11]; 3) The calling behavior of a spammer remains the same across many SPs; and 4) the detection approaches based on collaboration are more likely to have better detection accuracy and detection time than that of stand-alone detection approaches. Based on these assumptions, the rest of the chapter discusses a system called COSDS, which blocks spam subscriber based on the collaboration from the autonomous SPs. The basic components of our collaborative spam detection system are shown in a Figure 5.2. In the following sections, we describe the method used for computing *GR* of the subscriber (Section 5.4.2), the method used for classifying subscriber as spammer and non-spammer

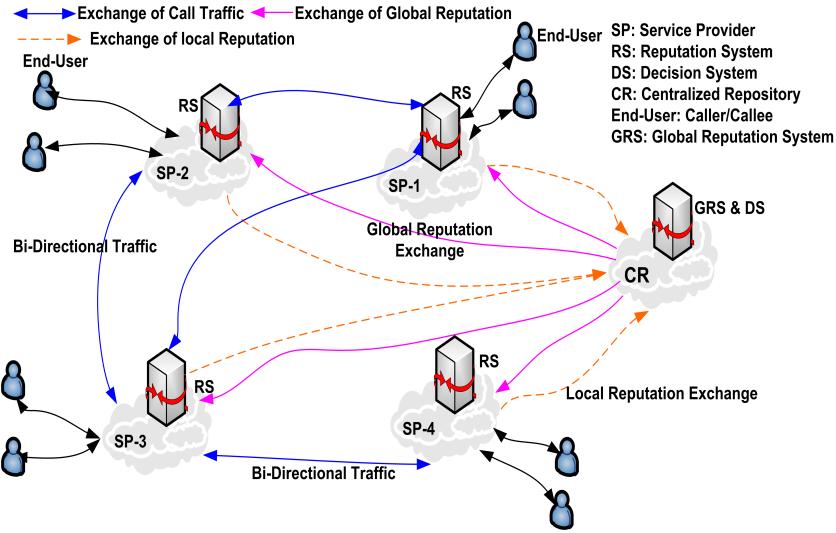


Figure 5.2: Building Block of Collaborative SPIT Detection.

(Section 5.4.3), and design options for collaboration and their effects on the detection accuracy (Section 5.4.4).

#### 5.4.1 System Design Overview

Figure 5.2 presents the system architecture of the COSDS system. COSDS consist of three layers. At the lowest layers, subscribers (end-users or subscribers) make and receive calls among each other. There is a local reputation engine placed in each SPs network that computes local reputation of a subscriber using his local call patterns extracted from the recorded CDRs at the SP. The SP reports *LR* scores to the trusted centralized repository in a following format: [CallerID, LR, Trust for SP]. CallerID is the unique identity of a subscriber (can be a telephone number, an IP address or both). We are using telephone number as the identity of the subscriber. The *LR* is the local reputation score of the subscriber and takes value in between 0 and 1. The third argument is optional and represents SP trusts score on other SPs from where it receives traffic or sends traffic to.

The *CR* is a trusted third party or regulator entity responsible for reputation aggregation and ensures that SP's provided information would not be disclosed to any other entity. The *CR* computes global reputation of subscriber by aggregating received *LR*, makes decisions about subscriber (spammer and non-spammer) and report back results to collaborating SPs in a following format [Caller ID, GR, Decision]. The *GR* score is the aggregated reputation of the subscriber and the decision is the status of a subscriber as a spammer and a non-spammer.

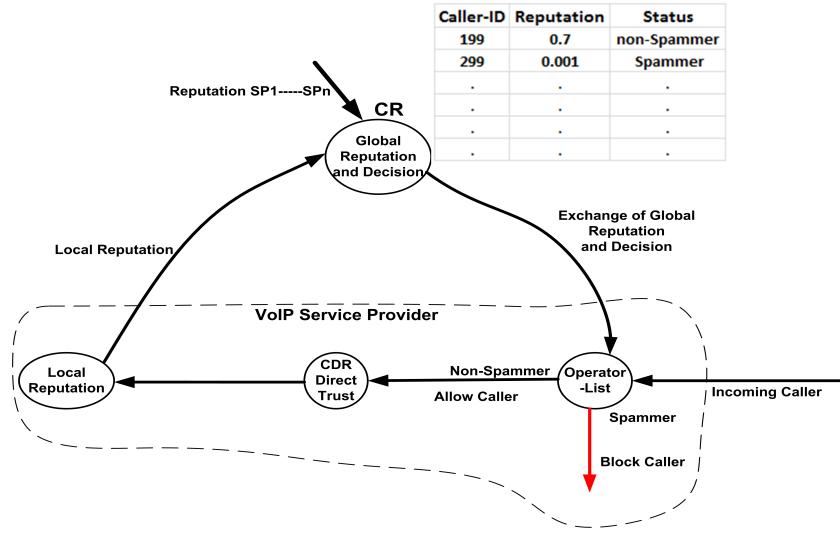


Figure 5.3: SP's Level Working of Collaborative SPIT Detection.

Figure 5.3 presents SP's reaction towards a call request from the subscriber. Upon reception of a call request, SP first checks caller-id against its local database. If a caller-id is present in a spam list then the SP immediately blocks the subscriber and if the subscriber is legitimate then SP allows subscriber to have communication with the callee. At the end of conversation, the SP updates direct trust between subscriber and callee, computes subscriber's *LR* and periodically sends *LR* to the CR. The CR computes *GR* of the subscriber considering new reputation score reported from the collaborators. At the end of a collaboration process, the SP receives aggregate *GR* and classification result of a subscriber from the CR. The SP can either rely on CR decision to update his spammer-legitimate database or uses *GR* scores along with social behavior of subscriber within the SP.

#### 5.4.2 Global Reputation of a Subscriber

The *GR* of a subscriber is computed in two steps. First, a SP computes *LR* of the subscriber and sends it to the CR, and secondly, a CR computes *GR* of the subscriber by aggregating received reputation scores from the collaborating SPs.

The computation of *LR* of the subscriber in a SP is a two steps approach. First, a direct trust between a subscriber and his called callees is computed from the subscriber's past call transactions with his callees. Second, a *LR* of the subscriber is computed using the Eigen Trust algorithm. Existing methods used for computing direct trust of subscriber uses two main approaches: getting positive and negative feedback from the callee about the subscriber and using call features from the CDR e.g. average call duration. Relying on

a subscriber's feedback is intrusive and relying on a average call duration will lead to allowing spammers having few good duration out-going calls from a large number of called recipients. A combined use of several features would be more effective in characterizing the real behavior of the subscriber within the service provider. In a combined approach, a direct trust between subscriber and his called callee is computed by collectively considering the in-coming and out-going call rate of the subscriber to the callee, the call duration of calls made and received between subscriber and the callee and the number of unique callees a subscriber has for a particular time window. These features have been adopted because of the fact that legitimate and spam subscribers exhibit different calling behavior. The legitimate subscribers usually have long duration, bi-directional repetitive calling behavior with their friends and family members, and also have small duration bidirectional calls to very few called callees. On the other hand, the spammer or the advertiser usually targets large number of callees, which normally results in a short duration calls to a large number of callees. Spammer also manages a moderate duration calls with a few target callees as well as receives calls from the few targeted callees. This unbalanced calling behavior of spammer i.e. high number of out-going calls and few in-coming calls would result in a small direct trust score for a spammer with the large number of called callees. Within a service provider  $SP$ , the direct trust  $Trust_{SR}^{SP}$  between subscriber  $S$  and his callee  $R$  is computed by using equation 5.1.

$$Trust_{SR}^{SP} = \frac{CD_{SR}^{SP} \times CallRate_{SR}^{SP} + CD_{RS}^{SP} \times CallRate_{RS}^{SP}}{PO_S^{SP}} \quad (5.1)$$

In equation 5.1,  $CD$  is the in and out call duration between a subscriber and the callee in a specific time interval,  $Call - Rate$  is a frequency of in and out calls made between subscriber and his callee in a specific time interval, and  $PO$  is the out-degree of the subscriber. The SP defines a sparse trust matrix of dimensions  $N \times N$ , where  $N$  is the total number of subscribers within the SP. If there is no interaction between subscribers then  $Trust_{SR}^{SP}$  from subscriber  $S$  to subscriber  $R$  is set to be zero. The direct trust between subscribers is asymmetric as the subscriber and his callee might have different number of out-going callees. For the spammers, equation 5.1 would result in a weak trust relationship with a large number of callees and moderate trust with only few callees. For the legitimate subscribers this would result in a strong trust relationship with many of his callees and moderate trust relationship with a large number of called callees.

**Algorithm 5.1** Global Reputation of Subscriber S At CR

---

```

1: procedure AGGREGATING REPUTATION OF SUBSCRIBER S ()
2:   input $\leftarrow$ Trust Matrix  $Trust_{SR}$  of Subscriber S With Callee R within SP Using Eq.1.
3:   output  $\leftarrow$  Global Reputation  $GR_S$  of the Subscriber.
4:   for for each SPs do
5:     for for All Subscriber within the SP do
6:       initially  $GR_S=1/PO_S$ 
7:       ;Iterate until convergence
8:       while  $\delta < \varepsilon$  do
9:          $LR_S \leftarrow Trust_{SR} \times GR_S$ 
10:         $GR_S \leftarrow LR / \|LR\|$ 
11:         $gr \leftarrow \|LR\|$ 
12:         $\delta \leftarrow \left| \frac{gr - gr_{previous}}{gr} \right|$ 
13:         $gr_{previous} \leftarrow gr$ 
14:      end while
15:    end for
16:  end for
17:  Send the LR Scores [ $CallerID$ ,  $LR_S$ ,  $Trust$  for SP] to the CR.
18:  for All subscriber S do
19:     $GR_S = \frac{\sum_{SP=1}^N W_{SP} \times LR_S^{SP}}{N}$ 
20:  end for
21:  Exchange of Global Reputation  $GR_S$  of a subscriber to each Collaborating SP.
22: end procedure

```

---

The SP then computes the  $LR$  of the subscriber by applying Eigen Trust algorithm to the normalized direct trust of score of the subscriber. The direct trust of a subscriber is normalized as follows:  $Trust_{SR}^{SP} = Trust_{SR}^{SP} / \sum Trust_S^{SP}$ . The subscriber's reputation in the SP is represented as  $LR_S^{SP}$  and is computed iteratively from  $LR_S^{SP} = Trust_{SR}^{SP} \times GR_S$ . Where  $GR_S$  represents the  $GR$  of the subscriber after collaboration. Initially, the  $GR_S$  of a subscriber is set equal to  $1/PO_S^{SP}$ . The computation of  $LR$  of a subscriber is an iterative process that continues until the average relative error is less than  $\varepsilon$  as shown in algorithm 1(lines 8-16). On each aggregation cycle the SPs sends  $LR$  ( $LR_S^{SP}$ ) of a subscriber to the  $CR$  and receives the aggregated  $GR(GR_S)$  of the subscriber from the  $CR$ .

The  $CR$  aggregates  $LR$  of the subscriber provided by the cooperating SPs and makes sure that reputation aggregation is done once for every reputation aggregation cycle.  $CR$  computes  $GR$  of a subscriber using weighted average algorithm applied to received  $LR$  scores and trustworthiness of cooperating SPs. The  $GR$  of the subscriber is computed as presented in equation 5.2.

$$GR_S = \frac{\sum_{O=1}^N W_{SP} \times LR_S^{SP}}{N} \quad (5.2)$$

In equation 5.2,  $N$  is the total number of SPs participating in a collaboration,  $W_{SP}$  is

the trust of a SP reporting reputation scores and  $LR_S^{SP}$  is the  $LR$  of the subscriber in a SP. The weighted average aggregation allows CR to assign different importance to different collaborating SPs. The spammers are not necessarily spamming all SPs equally and at the same time. They make spam calls to one SP and when blocked by the SP they choose recipients from another SP. The lower the  $LR$  of the subscriber in several SPs, the lower the  $GR$  is - despite the subscriber having good reputation in only few SPs. In other words, if the majority of the SPs assign small reputation value to subscriber  $S$  then subscriber  $S$   $GR$  would bend towards the reputation of  $S$  in the majority of SPs. The weighted aggregation method is suitable for  $GR$  aggregation and would assign a high reputation score to the subscriber only if the subscriber has high reputation in the majority of the called SPs. It is also necessary to assign different weights to the collaborating SPs. The call-receiving SP can assign weights to sender SPs on the basis of the fraction of subscribers classified as spammer to the total number of subscriber making calls to the recipients of SP.

$$W_{ij} = 1 - \frac{\text{No.of Subscribers from } SP_{ji} \text{ identified as spammers}}{\text{Total No. of Unique Subscriber from } SP_{ji}} \quad (5.3)$$

In equation 5.3,  $SP_i$  receives calls from  $SP_j$  and  $W_{ij}$  is the trust weight of  $SP_i$  for the  $SP_j$ .

The SP periodically exchanges the reputation score of their subscriber to the CR. The convergence process measures the number of cycles required for local and global convergence and affects the resources required for the collaboration process. In algorithm 1, the convergence is measured as the number of iteration required in step 8. This is bounded by  $\left| \frac{gr - gr_{previous}}{gr} \right|$  rate, where  $gr$  and  $gr_{previous}$  are the first and second largest eigenvalues of the trust matrix  $LR_S$ . In COSDS, convergence is carried out locally and final  $LR$  is sent to the CR for  $GR$  computation. This process is not resource intensive as each SP requires only two transmission cycles for computing  $GR$  of the subscriber; one transmission cycle for sending its  $LR$  to the CR, and one cycle for receiving the  $GR$  from the CR. The communication overhead is much less than the communication overhead required for computing  $GR$  by having direct collaboration among SPs in a distributed way similar to how P2P trust systems work. COSDS is also independent of the number of pre-trusted subscribers and the traffic overhead remains constant even if the out-degree of the subscriber and number of collaborators increases but depends on the number of subscribers within the SP.

### **5.4.3 Detection of SPIT Subscriber**

CR maintains a vector representing the  $GR$  of each subscriber and has a value between 0 and 1. The subscriber can be classified as spammer if  $GR$  of a subscriber is below a threshold value  $T$ . There are two design choices for selecting the  $T$ : 1) a fixed threshold

based on a TP or FP tolerance policy of each SP, and 2) a dynamic threshold chosen automatically from the global reputation of the subscribers. The design of a fix threshold is straight forward, but it does not necessarily adopted with the calling behavior of subscribers. Computing the dynamic threshold requires analysis of present and past calling behavior of all subscribers and is adoptive.

In COSDS, we expect that subscriber having small  $GR$  scores are more likely to be considered as spammers and subscribers having high  $GR$  score are likelier to be consider as the legitimate. Spammers usually have similar call patterns and their  $GR$  lies near to each other and much distanced from the  $GR$  scores of the legitimate subscribers. In addition, SP may also wish to block the top spammers. Moreover, new technologies normally witness a very few malicious subscribers until they become mature and attract large number of subscribers. However, once it attracted large number of subscribers, it start attracting more malicious subscribers and the percentage of spammers rises to as up as 40% of all identities joined the network on a particular day. For example, currently, it is estimated that 36% tweets on a tweeter contains spamming links [TWI16] and 25% of all personal computers may be infected by Viruses. Over the time, we expect similar behavior in case when telephony becomes less costly and primary method for the personal and business communication.

Considering these facts and intension for blocking only top spammers, we adopted a percentile-based dynamic threshold [AM13] approach for computing the classification threshold below which subscriber is classified as spammer. Specifically, we compute the distributions for the reputation scores for two classes: spammer and non-spammers. The procedure for classifying subscriber as a spammer or a non-spammer is presented in algorithm 2. In algorithm 2, first, the 25th percentile of  $GR$ ( $GR$  vector) is computed and then the mean  $m$  of the  $GR$  score of all subscriber below the 25th percentile is used as a final threshold  $T$ . Subscribers can be classified as spammer 1 or non-spammer -1 based on a following rule:

$$\text{Subscribers}_S = \begin{cases} GR_S > \beta \times \text{threshold} & ; 1 \\ GR_S < \beta \times \text{threshold} & ; -1 \end{cases}$$

Spammers normally increase their number of callees over time. Although, it might be possible that spammers bypass our system during first few aggregation cycles but over time or after few iterations they would not be able to bypass COSDS system. Specifically, COSDS would have almost zero FP rate (non-spammer classified as spammer) under high

**Algorithm 5.2** Detecting Spammer and Updating Service provider Trust
 

---

```

1: procedure SPIT SUBSCRIBER ()
2:   input $\leftarrow$ Reputation  $GR_S$  and threshold  $\beta$ 
3:    $SP-defined\ parameter \leftarrow \beta$  ( $\beta = 1$  if  $SP$  has no preference)
4:    $m \leftarrow 1st-quartile(GR_S)$ 
5:    $T \leftarrow mean(GR < m)$ 
6:   for All subscriber  $S$  do
7:     if ( $GR[S] < \beta \times T$ ) then
8:       Subscriber  $S$  is Spammer
9:     else
10:      Subscriber  $S$  is non-Spammer
11:    end if
12:   end for
13:   Update Weights of  $SP$  using Eq.3
14: end procedure
  
```

---

spamming rate, thus maximizing SPs profit by not blocking legitimate subscribers. The 25th percentile would not provide optimum detection when percentage of spammer exceeds 30%. A small adjustment in the threshold could improve detection performance even under when number of spammers are very high. The threshold  $T$  can also be implemented with a SP's parameter  $\beta$  (greater or less than 1) in order to behave under extreme conditions i.e. extremely high and low spamming rates. The CR responds to the SP with a vector representing  $GR$  and status of all subscribers. The SP can also adopt different threshold based on its own detection tolerance or use local call and social features such as the inter-arrival time of the call request, the call rate along with  $GR$  for the final classification. The SP can also ask callee to accept or reject the call by sending  $GR$  score of the subscriber calling to the callee at the time of call request but this requires changes in a call request message and is intrusive to the callee.

#### **5.4.4 Design Options for Information Summarization and Collaboration**

VoIP SP records subscribers call transactions in a CDR database that holds information about each call made or received by the subscriber of the SP. CDR contains diverse set of information including caller-callee unique identifiers (calling identities), IP addresses of the subscriber involved in communication, duration and time of a call, and a call status (successful, failed, busy).

A very high accuracy (correct classification of spammers as a spammers and non-spammers as non-spammers) is expected, if the SP exchanges CDR or direct trust scores of subscribers to the CR but does not ensure privacy. Moreover, the exchange of CDR

and direct trust also increases the communication overhead. The major challenge in design of collaborative system is to achieve a trade-off between information summarization, detection accuracy and communication overhead.

The trade-off between accuracy and information summarization can be achieved in three levels: 1) no summarization, 2) call detail summarization, 3) network summarization.

In a first design option, the SP filters locally recorded CDRs and exchanges CDRs containing caller-callee identities, call duration, and call time to the CR. The CR aggregates the CDRs from all collaborating SPs and computes *GR* of the subscriber using approach presented in chapter 4 . This means that a large amount of data needs to be transmitted to and processed at the CR. Although the trusted CR guarantees protection of sensitive information provided by the SP but an intruder on the CR is still able to infer social network of a target subscriber. This design option may provide better detection accuracy because of the availability of complete information about behavior of subscriber from all SPs on a single place but has limitation of communication overhead. This design choice also increases computation load on the CR because of processing of millions of call records. Another limitation of this approach is that SPs are not easily convinced to provide raw CDRs of their customers to trusted and non-trusted authorities.

In a second design option, the SP sends caller-callee direct trust scores to the CR. In this case, the SP computes direct trust between subscribers from locally recorded CDR by using equation 5.1 and sends direct trust to the CR. The CR then aggregates direct trust scores, computes *GR* of subscribers and classify them as a spammers or a non-spammers using approach presented in chapter 4. This design option has benefit of distributing computation load between SPs and the CR, and hides subscriber critical personal information such as call-rate and call duration. However, intruders at CR are still able to infer the social relationship information and trust of subscriber with his called callees from the trust scores.

In a third design option, the SPs locally compute the *LR* of a subscriber within the SP and exchange this score to the CR. The CR then computes *GR* of the subscriber by aggregating *LR* scores and make decision about subscriber based on an aggregated *GR*. This approach has a small communication overhead compared to other collaboration options and fully protects privacy of subscribers.

It is expected that CDR based collaboration has better detection accuracy than the COSDS approach. This is because in a reputation based collaboration, SP computes the *LR* of subscriber and has no information about subscriber behavior in other SPs. On the other hand in CDR based collaboration, the whole call-rate, call duration and out-degree

of subscriber across all called SPs is known to the CR for the computation of global reputation of a subscriber.

### **5.4.5 Communication overhead**

Each SP taking part in a collaboration process incurs a communication overhead. The communication overhead between a collaborator and the CR depends on the amount of information delivered from the collaborator to the CR and the number of communication rounds taking place between them. Our approach reduces the communication overhead by reducing the amount of information delivered from the SP to the CR and reduction of communication rounds between SP and the CR. In COSDS, the SP stores and delivers  $LR$  of his subscribers to the CR which requires only 22 bytes for a one subscriber (14 Bytes for Caller-ID and 8 Bytes for the Reputation score). The CR computes  $GR$  and makes decision about subscriber (spammer or non-spammer) and delivers this to the all collaborating SP which requires 23 bytes (14 bytes for the Caller-ID, 8 Bytes for the Global Reputation and 1 Byte for the decision). The overall communication overhead require for sending scores to CR is  $n * 22$  Bytes (where  $n$  is the total number of subscribers in a SP) and communication overhead requires for sending  $GR$  and decision to collaborating SP is  $k * 26$  Bytes per SP (where  $k$  is total number of subscribers from all SP). The exchange of CDRs and direct trust require high communication and memory overhead when compared to COSDS. The overhead requires for sending CDR to CR depends on the total number of call records and exchange of direct trust depends on a total number of subscribers in the SP.

### **5.4.6 Discussion on Privacy Protection**

VoIP SP needs to protect privacy of his customers in two aspects: 1) Protection of subscriber's Pseudonymized Identity: preventing the adversary having some auxiliary (AUX) information to find anonymized identity of his target; 2) Social Relationship Network Protection: The existence of social relationship and strength of social relationship between target subscriber and his friends should not be learned by the adversary.

#### **5.4.6.1 Adversary Model and the Privacy Breach**

We assume that subscribers are not intrusive and SPs are not misusing recorded CDRs but there exist some intruders who wish to extract private information of subscribers. The adversary has some information about the target subscriber, which may include time of few calls, call duration of few calls, and the call rate. The objective of adversary is to use this auxiliary information to learn pseudonymized identity and infer social relationships

of the target subscriber. The Probability that an adversary can breach the privacy and get true records given AUX information is presented as:

$$Pr(PrivacyBreach|AUX) = \begin{cases} 1/X & ; if X > 0 \\ 0 & ; if X = 0 \end{cases} \quad (5.5)$$

Where X is number of subscribers returned for the AUX information.

#### 5.4.6.2 Privacy Protection at SP

The SP processes CDRs for the computation of *LR* score of the subscriber. The adversary learns following information for breaching the privacy of subscriber during computation of *LR*:

**AUX1:** An adversary knows call related information of the target user and wants to find anonymized identity of the target user. For example, an adversary knows target user called someone known person at 11:20 am.

**AUX2:** An adversary knows out-degree of target user along with AUX1. For example, an adversary knows call times of calls made by the target user and number of callees target user called.

**AUX3:** An adversary knows the calling behavior of target user along with AUX1 and AUX2. For example, an adversary knows call rate and call duration of target user's few calls and wants to learn complete relationship network of the target user.

COSDS protects privacy of the subscriber within the SP by setting the following best practices. 1) The SP shall protect records of the subscribers from unauthorized access using strong authentication processes, 2) The SP shall provide opt-op option to the subscriber if his out-degree is small, and 3) The SP shall pseudonymized identity of subscriber for further reducing the risk of misuse of the data. Pseudonymized identities can provide one level of protection but adversary can still find pseudonymized identity of target by using single AUX or correlating multiple AUX. In Section 5.6.7, we will show that Pseudonymized identity is not providing absolute privacy protection for AUX1, AUX2 and AUX3. We use following mechanism for the CDR anonymization:

**P1:** The local reputation engine computes reputation for the specific time period, we strip the minutes and seconds information from the time and date of the CDR. By doing this the probability of inferring the pseudonymized identity is small for the AUX1.

	Social Network	Calling Behavior	Trust Network
Complete CDR	YES	YES	YES
IDs with Rate and Duration	YES	YES	YES
IDs with Trust Matrix	YES	NO	YES
IDs with Reputation Scores	NO	NO	NO

Table 5.1: Subscriber Level Privacy Breach for Different Collaboration Methods.

**P2:** The out-degree of the subscribers in the CDR can be k-anonymized. For subscribers having unique out-degree, random noisy subscriber can be generated which is exactly similar to the subscriber but with different pseudo identity. This k-anonymization would affect the detection accuracy but provides privacy protection for AUX2.

The adversary knows AUX1 of his target subscriber; for example, adversary learns from media that presidents of two countries talk to each other for some duration on some specific time and wants to learn pseudo identities associated with both presidents. The adversary can find possibly a small candidate-set if time in CDR is not properly anonymized and by correlating more information adversary can find correct identities of both presidents. However, in our scheme, striping minutes and second further minimizes the risk of de-identification. In some scenario, the adversary can make some calls to the target subscriber and intends to find whether the target has interactions with his friends or not. In a first case, adversary knows call duration and call time of all his calls to the target subscriber. Again, the adversary can learn his target's pseudonymized identities and so the presence of link between target and his other friends if time of call is not stripped. The adversary can also correlate multiple AUX to reduce the size of candidate set. However, our proposed anonymization approach significantly reduces the risk but adversary can breach privacy by making some large number of bi-directional links which are normally not under his control.

	Subscriber Home SP	Calling Behavior	SP Network
Complete CDR	YES	YES	YES
IDs with Rate and Duration	YES	YES	YES
IDs with Trust Matrix	YES	YES	YES
IDs with Reputation Scores	YES	NO	NO

Table 5.2: Service Provider's Level Privacy Breach for Different Collaboration Methods.

### 5.4.6.3 Privacy Protection at CR

In a centralized collaboration, the *CR* computes *GR* of subscribers by aggregating information received from the collaborating SP. The use of trusted CR ensures that provided information would not be misused but still has possibility of privacy breach attack by the adversary. The exchange of reputation scores to the trusted CR is not disclosing any information about underlying relationship network of subscribers but adversary or other SP can try to infer some information about target given local and *GR* scores. The goal of adversary at a CR is to utilize the reputation of the target subscriber and learn his possible relationship network. In some scenarios, the SP itself become adversary and wants to learn relationship network of target belonging to other SP from the received *GR* and locally recorded CDR of the target. The adversary has the following AUX information at CR:

**AUX4:** The adversary knows *LR* of target user and other subscribers in a target SP. The adversary also learns that target user only interacts with highly reputed subscribers or subscribers having similar reputation scores. The goal of adversary is to predict possible relationship network of the target user in a target SP.

We assume that the communication between CR and collaborating SPs is secure. The exchange of single reputation score ensures privacy protection against AUX 1, 2 and 3 as shown in Tables 5.1 and 5.2 for subscriber level and SP level privacy breach. However, the adversary SP can make a guess about relationship network of target subscriber given AUX 4 but the probability of breach is extremely small and further computationally impossible when the number of reputed subscribers are high.

## 5.5 Experimental Methodology

In this section, we provide an overview of method used for generating the synthetic data-set and the evaluation criteria used for evaluating the performance of proposed detection system.

### 5.5.1 Synthetic Data-Set

We generated a synthetic call detail records using same approach as discussed in chapter 4. Our objective is to generate synthetic CDRs that exhibit similar characteristics to that of the real world CDRs. The communication behavior of the subscriber within the service provider is modeled through three fundamental features such as call-rate, call

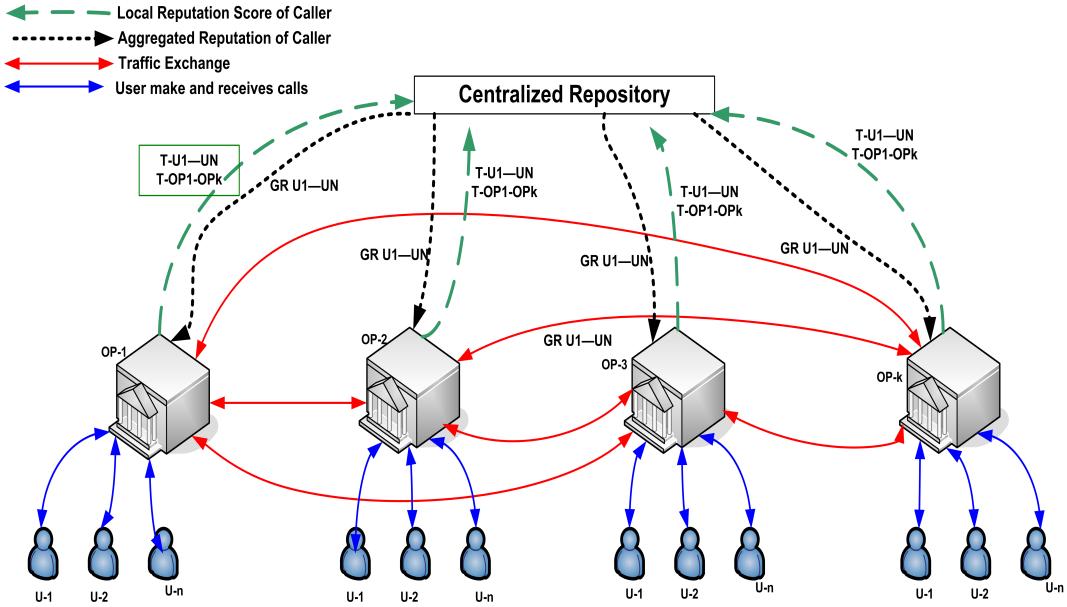


Figure 5.4: Collaborative Simulation Model.

duration and number of unique callees of the subscriber. The communication behavior of legitimate subscriber is different from the spammer in perspective of following three features. Firstly, the legitimate subscriber normally has long duration calls with many of his recipients, whereas SPIT subscriber has large number of short duration calls with his called recipients because recipient disconnects call immediately as soon as he realized that subscriber is a telemarketer or an advertiser [CMP+13], [DTN11]. Secondly, the SPIT subscriber does not exhibit repetitive calling behavior, but legitimate subscribers have repetitive calling behavior with many of his callees [DTN11]. Thirdly, the SPIT subscriber targets large number of recipients thus has high out-degree, whereas the legitimate subscriber has some controlled out-degree.

The basis of our synthetic CDRs is a graph representing the social network of the subscribers in a service provider. Vertices represent the calling identity of the subscriber and edges between vertices represent the phone call between two vertices (caller and the callee). In a simulation setup, legitimate subscribers has following distributions for the out-degree, call duration and call rate [AM13]. 1) The degree of legitimate subscriber fits into the power-law distribution [NGD+06]. In order to have a power-law distribution, we modeled the social network of subscriber as a Barabasi-Albert graph model with the average out-degree of 10 [NGD+06]. 2) The call duration of the legitimate subscriber is modeled using exponential distribution with the average call duration of 360 seconds. 3) The call rate of the legitimate subscriber is modeled using Poisson distribution with mean value of 5 calls. The simulation model in Figure 5.4 consists of five VoIP service providers and each service provider has 50000 legitimate callers with different percent-

ages of spammers. The callees of legitimate caller are distributed across all the service providers with 60% of the callees belong to callers registered network and remaining 40% are equally distributed in across other service providers.

The spammer usually tries large number of callee. In simulation, spammer calls 10%-to 30% of unique calls per day and each callee is randomly selected from the legitimate caller. The call duration of the SPIT subscriber is modeled through exponential distribution with different average duration i.e. average duration of 180 with few callees and average duration of 60 seconds with majority of callees [CMP+13], [DTN11]. The degree of the SPIT subscriber is randomly chosen between 500 and 4000 [CMP+13] and has non-repetitive calling behavior. The spammer equally distributes callee across all the service providers.

We provided results for 2, 3, 4, and 5 collaborators. Each SP consists of 50 thousands legitimate and different percentage of spammers. The collaborating SP computes *LR* of their subscribers and periodically updates CR with subscriber's reputation. In a simulation, the SP updates CR after one day and all SPs update CR at the same time. We assume secure communication channel between SP and the CR for the exchange of information. For each scenario, we performed simulations for 10 times and show the average results with standard deviation.

### 5.5.2 Evaluation Metrics

We use the standard information retrieval metrics of True Positive (TP) rate, False Positive (FP) rate and Accuracy (ACC) to measure the spam detection capability of COSDS. The true positive rate is defined as the ratio of the number of subscriber identifies as spammer to the total number of spammers. A legitimate subscriber that is classified as a spam subscriber by the detection system is termed as a false positive. The false positive rate is defined as the ratio of the number of false positive subscriber to the total number of legitimate subscriber in the call record. The evaluation metrics can be explained through the confusion matrix illustrated in a Table 4.1. The TPR, FPR and accuracy is computed as  $TPR = TP/(TP+FN)$ ,  $FPR = (FP)/(TN+FP)$  and  $ACC = (TP+TN)/(TP+TN+FN+FP)$ .

## 5.6 Performance Evaluation

In this section, we present the performance results of COSDS and compare its performance to the performance of Caller-REP and Call-Rank. Additionally, we also provide privacy breach analysis for the different auxiliary information.

### 5.6.1 True Positive Rate

We evaluated detection rate of COSDS and other system for three parameters: TP rate over the time, TP rate against different percentages of spammers, and TP rate against varying number of collaborators. It can be seen from a Figure 5.5 that COSDS approach out-performs other approaches and is able to block almost all spammers within 3 days in any percentage of spammers. Specifically, COSDS manages to achieve a TP rate greater than 80% on a first day which increases further to a 100% TP rate over the time regardless of number of spammers in the network. On the other hand, the non-collaborative Caller-REP achieves TP rate of up to 97% in 5 days when the number of spammers are small and prolongs detection time when the number of spammers in the network are high. This behavior is due to the fact that stand-alone detection systems only consider the local view of subscriber while computing direct trust and reputation of the subscriber within the SP. Nonetheless, non-collaborative systems are still capable of identifying local spammers and spammers from the other service providers spamming at a high rate. The improved performance of COSDS is attributed to the followings: firstly, it collectively uses different features while computing local reputation of the subscriber within the service provider; and secondly, SP collaborates for the computation of global reputation. The TP rate of COSDS increases as the number of collaborators increases since more collaborators are providing information about subscriber's reputation in their network as shown in Figure 5.5. Figure 5.5 also reveals that TP rate increases over the time regardless of number of spammers and it can also be seen that COSDS achieves almost similar detection rate to that of CDR based collaborative system.

From a Figure 5.5, we also observe that the TP rate of COSDS decreases slightly with the increase in the number of spammers and decreases considerably more for the non-collaborative system. Specifically, the TP rate of COSDS decreases to 60% when percentage of spammers varies from 40% to 70% as shown in a Figure 5.9.A. This can be further improved by using SP's-defined  $\beta$  parameter greater than 1.

Figure 5.5 also presents the detection rate of Caller-REP and Call-Rank under varying number of spammers. The detection rate of COSDS is much better than that of Caller-REP and Call-Rank. We observe that Call-Rank has degraded detection rate when compared to the Caller-REP. We can attribute TP rate of Caller-REP and Call-Rank to the following. Call-Rank considers average call duration while computing global reputation of the subscriber which allows spammer having some good duration calls to achieve high reputation scores within the network despite having high out-degree. On the other hand Caller-REP collectively uses call-rate of the caller in both directions, call duration of caller in both directions, and out-degree of the caller which results in a small reputation score to those

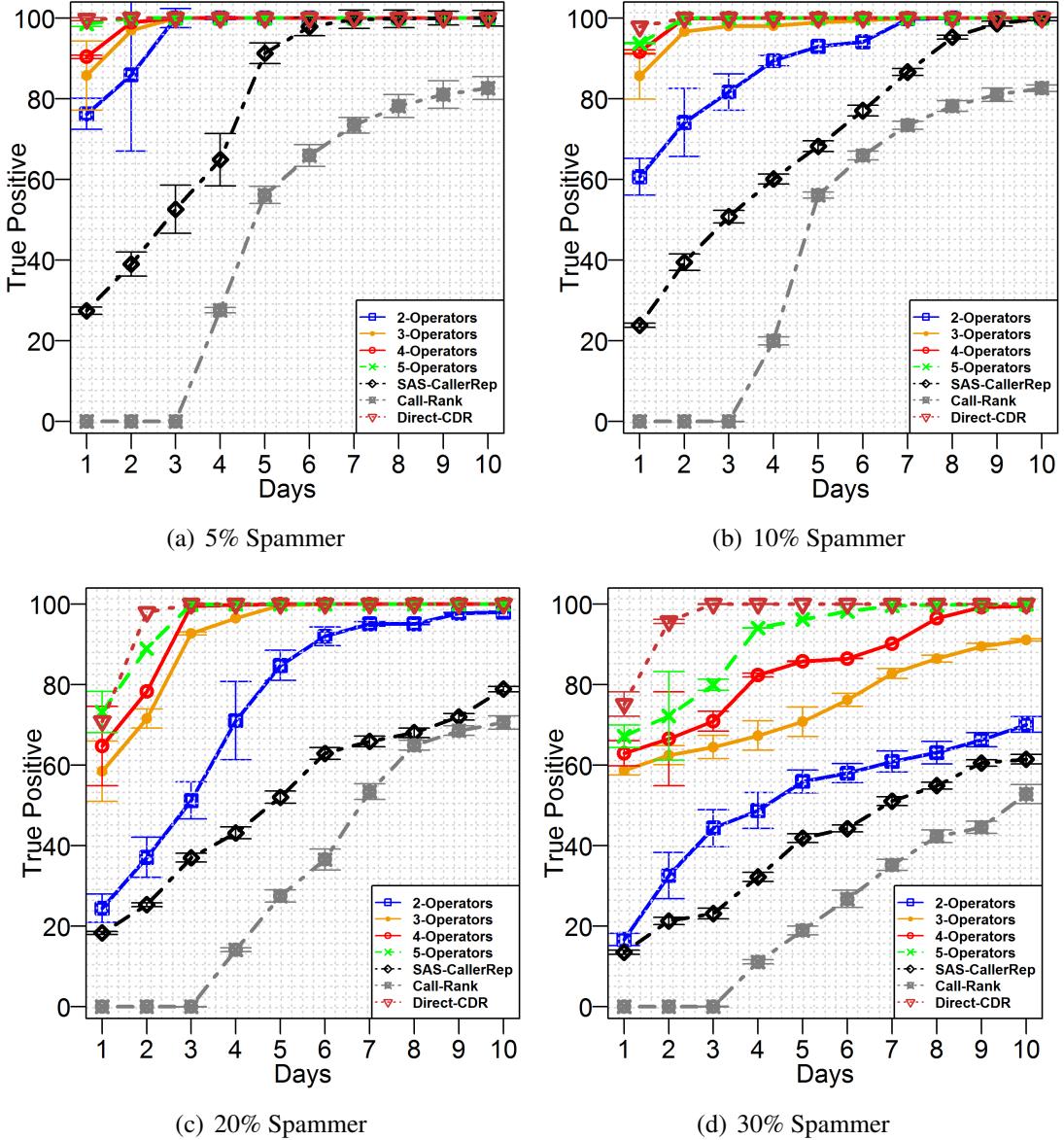


Figure 5.5: True Positive Rate of COSDS for SP trust=1 and  $\beta$  threshold=1.

subscribers having high out-degree and manage some long duration calls.

The scalability of the collaborative system is dependent on the number of collaborator participating in collaboration. The results from figure 5.5 show that 4 SP are enough for blocking above 98% of SPIT subscriber regardless of spamming rate. In a scenario where collaboration scores are received from the 50% of the total SPs, COSDS manages to block all spammers in 4 days for a spamming rate of less than 20% and achieves detection rate of more than 90% in 10 days when the number of spammers exceeds 20%. It can also be observed that the detection rate increases and the detection time decreases with the

number of collaborators.

We also observed that in all simulation scenarios COSDS approach provides almost - but not exactly - similar detection rate as the Direct-CDR approach. This happens despite COSDS preserving privacy and being less computationally and network demanding.

### **5.6.2 False Positive Rate**

Although TP rate is the key performance measure for evaluating the performance of SPIT detection system but it should have ideally zero FP rate. The FP rate not only annoys legitimate subscribers and callees but also results in a revenue loss for the SP because of blocking a legitimate subscribers. COSDS outperforms non-collaborative system in terms of FP rate and achieves FP rate of 0% in 3 days when percentage of spammer is high. The non-collaborative Caller-REP suffers from a high FP rate even after 5 days as shown in a Figure 5.6. The FP rate decreases further with the number of collaborators. It can be seen from a Figure 5.6 that with the 5 collaborators COSDS achieves FP rate less than 5% in three days for any percentage of spammers. Specifically, under a small percentage of spammers such as 5% and 10%, COSDS misclassifies large number of legitimate subscribers as spammers on a first few days and further improves it to FP rate of less than 5% within 3 days as shown in Figures 5.6.A and 5.6.B. In a condition of high spamming rate COSDS manges to achieve almost zero FP rate in 2 days with five collaborators as shown in Figures 5.6.C and 5.6.D. The FP rate of non-collaborative Caller-REP and Call-Rank is not acceptable as both have FP rate more than 5% even after 5 days. Specifically, the non-collaborative Caller-REP system has FP rate of more than 15% on first few days and achieves almost zero FP rate in 10 days which is too late.

COSDS uses local reputation scores for the computation of global reputation and decision about the subscriber. Some social behavioral features may provide some additional information about subscriber, for example a subscriber cannot be categorized as a spammer if his out-degree is extremely small. The FP rate can be further minimize by allowing SP to utilize other behavioral features along with the received global reputation scores and decision from the CR. Few such features are number of unique callees of the subscriber or ratio between total calls and number of friends. The FP rate can also be minimized by using a fixed threshold  $\beta$  defined by the SPs according to their requirements. The FP rate of COSDS and collaboration with Direct-CDR is almost similar to each other.

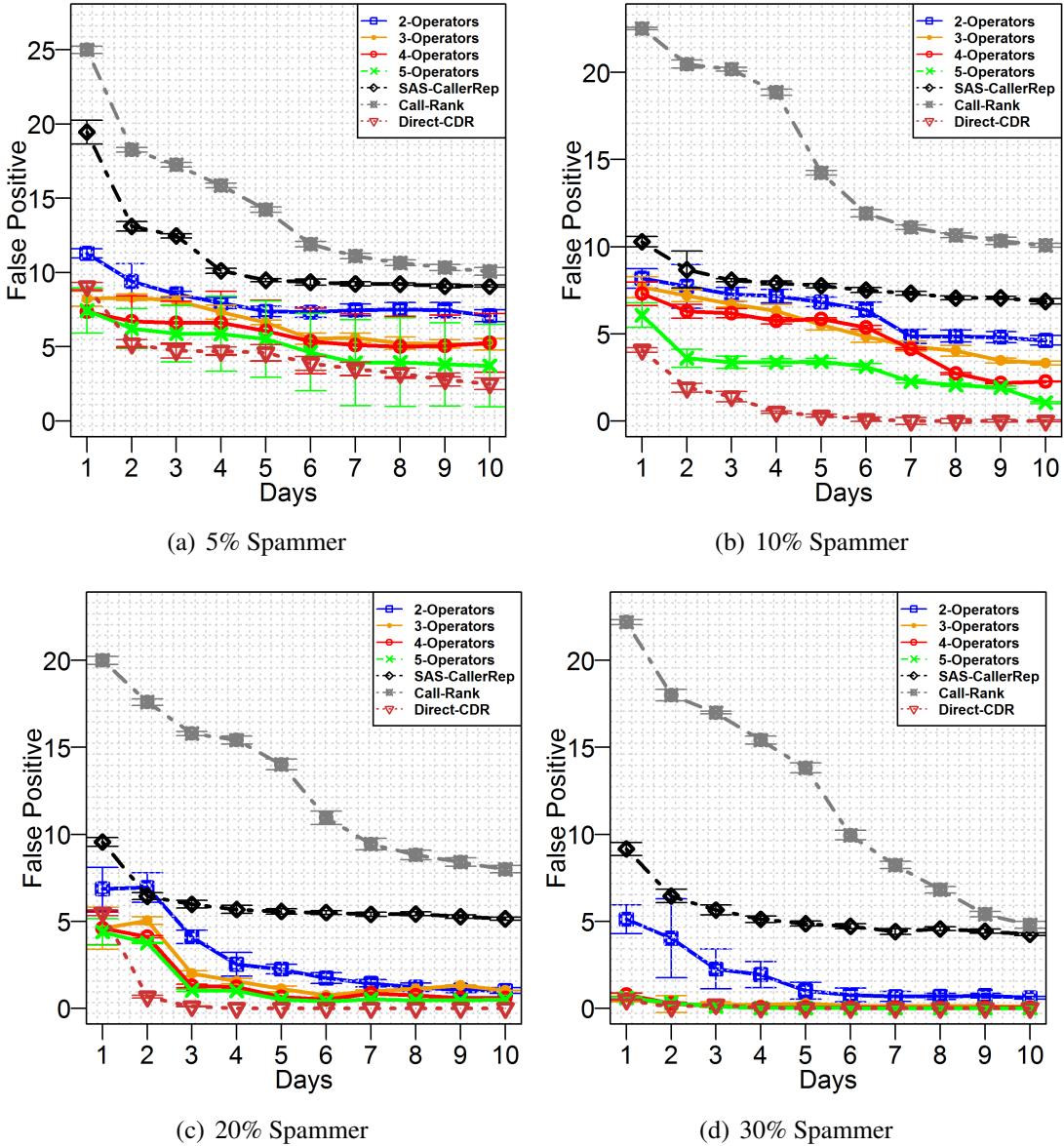


Figure 5.6: False Positive Rate of COSDS for SP trust=1 and  $\beta$  threshold=1.

### 5.6.3 Detection Accuracy

The detection accuracy is the proportion of true identification (both true positives and true negatives) to the total number of subscribers (either spammer or legitimate). It characterizes system's capability of making correct decision about all subscribers (classifying spammer as a spammer and non-spammer as a non-spammer). Under small spamming rate, the COSDS approach achieves high true positive rate with considerably high FP rate. However, under high spamming rate, COSDS manages to achieve better detection rate with a small FP rate. Figure 5.7 shows the detection accuracy of COSDS and other

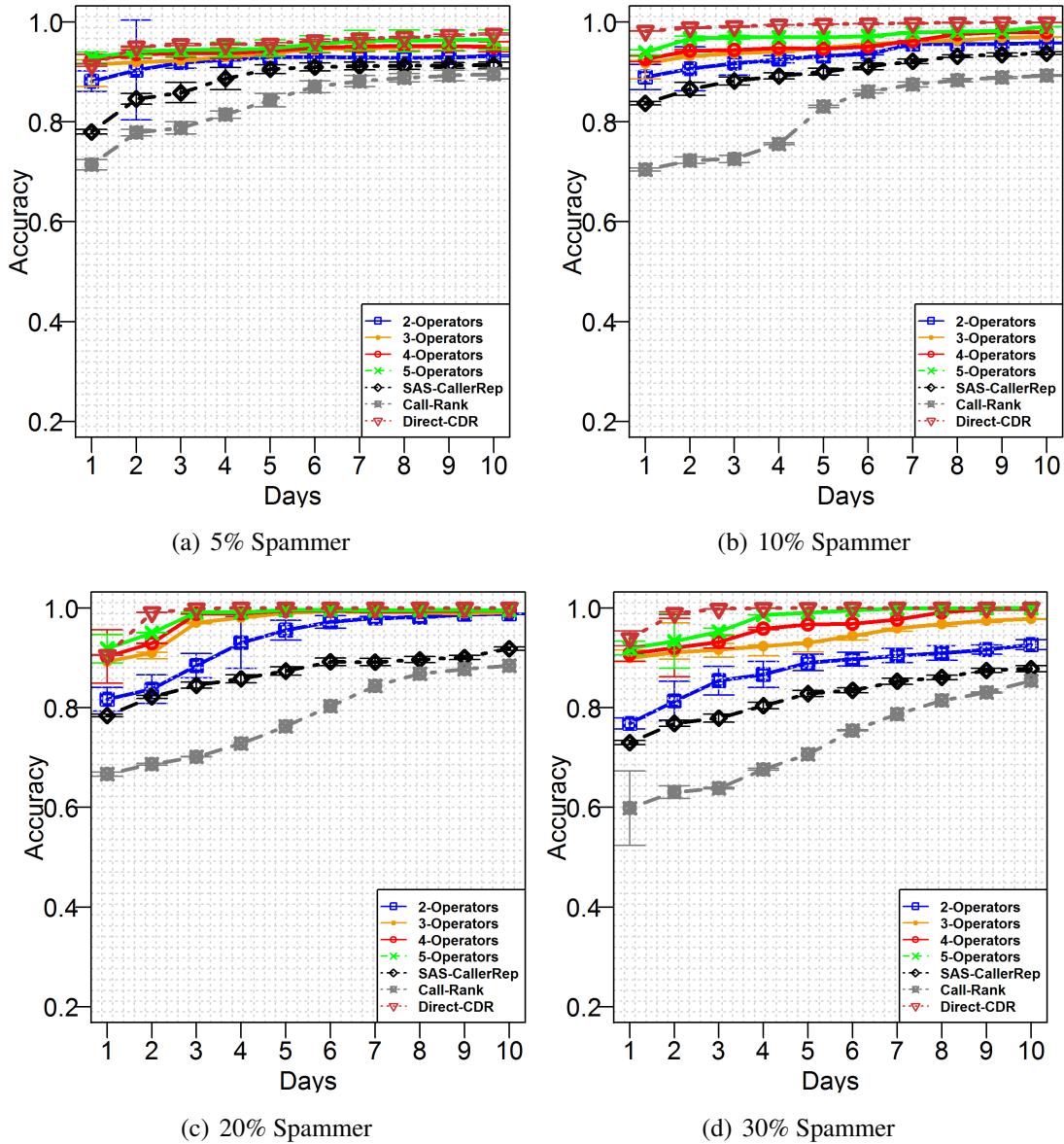


Figure 5.7: Detection Accuracy for COSDS and non-collaborative system for SP trust=1 and  $\beta$  threshold=1.

approaches when the number of spammers varied from 5% to 30%. COSDS achieves high detection accuracy than non-collaborative system because of collaboration which increases TP rate and decreases FP rate. Our experimental results show that, on an average, the accuracy of COSDS with five collaborator reaches to almost 100% in 4 days for any spamming rate which is much better than non-collaborative system as shown in a Figure 5.7. Specifically, we observe that COSDS reaches an overall accuracy of 99% in 5 days when the number of spammers are small ( $<10\%$ ) and reaches to overall accuracy of 99% in three days when number of spammers are high ( $>10\%$ ). This is due to the fact that

at small spamming rate COSDS misclassifies many legitimate subscribers as spammer i.e. about 7% in 3 days, and further goes to less than 2% in 5 days. CDR based collaboration indicates high detection accuracy than the COSDS approach. However, CDR based collaboration has some drawbacks in practical implementation as it requires more communication overheads, processing overheads and also exchanges private information of subscriber. The detection accuracy can be further improved by incorporating other social network features such as out-degree, clustering coefficient etc. along with the global and  $LR$  scores.

#### 5.6.4 Information Summarization and System Performance

The detection accuracy increases with the amount of information being exchanged but it requires system and network resources. The recommended system should provide high TP, small FP and high detection accuracy without too much system and network overheads. Figure 5.8 provides results for TP rate, FP rate and detection accuracy for different collaboration mechanisms.

As represented in Figure 5.8, the detection accuracy is opposite to that of information summarization. If SP does not participate in a collaboration process then no information is being exchanged and SP has own stand-alone detection system. The standalone system considers local view of the subscriber thus has poor detection accuracy on first few days. We considered three different collaborative mechanisms in our experiments: 1) Collaboration with complete CDRs; 2) Collaboration with direct trust scores of subscribers with their callees; and 3) collaboration with the reputation score of subscribers. The CDR based collaboration achieves best detection accuracy but it requires extensive network resources and also increases load on the CR. Additionally, the exchange of CDRs contains private information of service provider operational aspects and his customers. The exchange of direct trust summarizes the subscriber's call information but is not hiding subscriber's relationship network. Though it minimizes the load from the CR but still has a threat of privacy breach of subscriber. In comparison to first two approaches, COSDS collaborates with the summarized information without sending either CDR or relationship trust network to the CR. The results from Figure 5.8 show that the detection accuracy of collaboration with summarized information and collaboration with CDR become same over the time. The results in Figure 5.8 are shown for 5 collaborators and 30% spammers. The data used for the Figure 5.8 has been taken from Figures 5.5, Figure 5.6 and Figure 5.7.

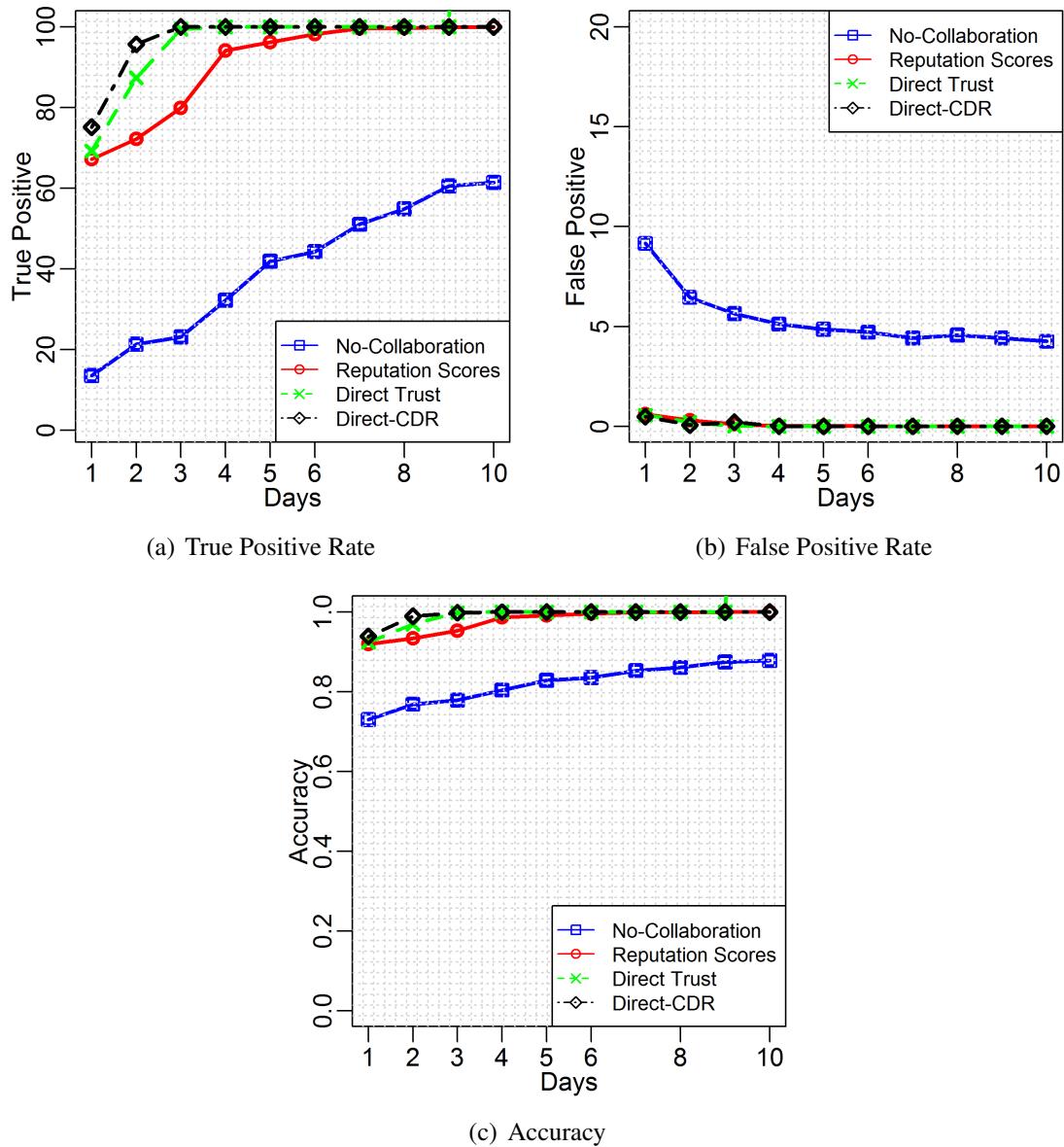


Figure 5.8: Information Summarization and True-Positive, False Positive and Accuracy trade-off for different Collaboration Methods.

### 5.6.5 Effect of Threshold on Performance

Until now, we have provided the performance results for only percentile based threshold. COSDS approach can also be implemented with the additional parameter defined by the service provider according to its defined policies for spammer and non-spammer. The service provider set value for the  $\beta$  parameter that is used along with the percentile threshold. The choice of an optimal  $\beta$  parameter depends on the relative tradeoff between the true positive rate and false positive rate. The major challenge in choosing a  $\beta$  parameter is

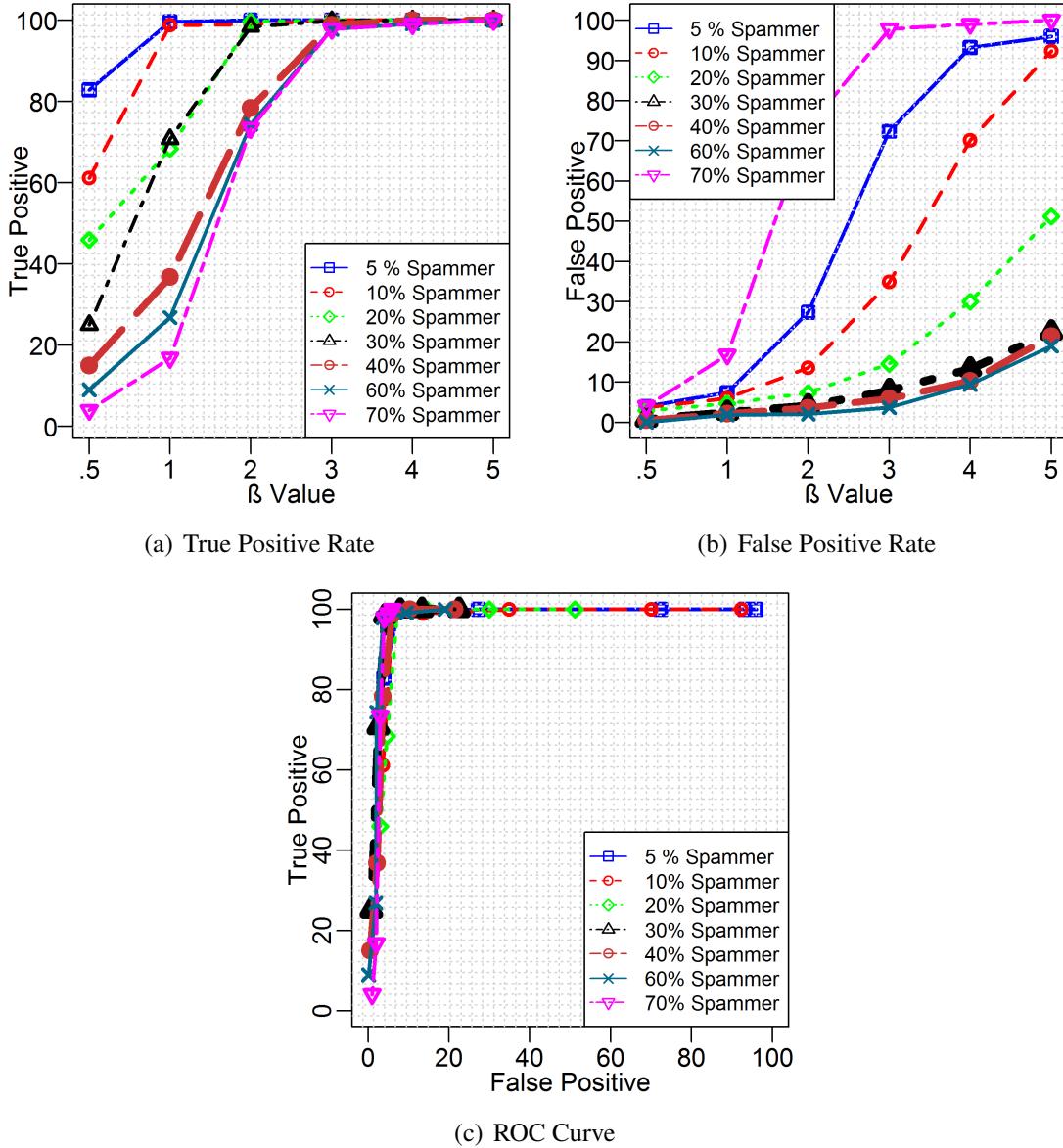


Figure 5.9: The Effect of Threshold  $\beta$  for TP and FP Rates for 5 Collaborators and for the First Day.

that it would incur small false positive rate and has high true positive rate. The TP and FP rate of COSDS for different  $\beta$  values and different percentage of spammers is shown in Figure 5.9.A and Figure 5.9.B. It can be seen from Figure 5.9.A that COSDS is not providing optimal TP rate at a small  $\beta$  value. Specifically, in a network with more than 30% spammers it shows poor resistance against spammers but provides better FP rate for any percentage of spammers as shown in Figure 5.9.B. The choice of high  $\beta$  value could improve TP rate to 100% for any spamming rate, but it has high FP rate for a network

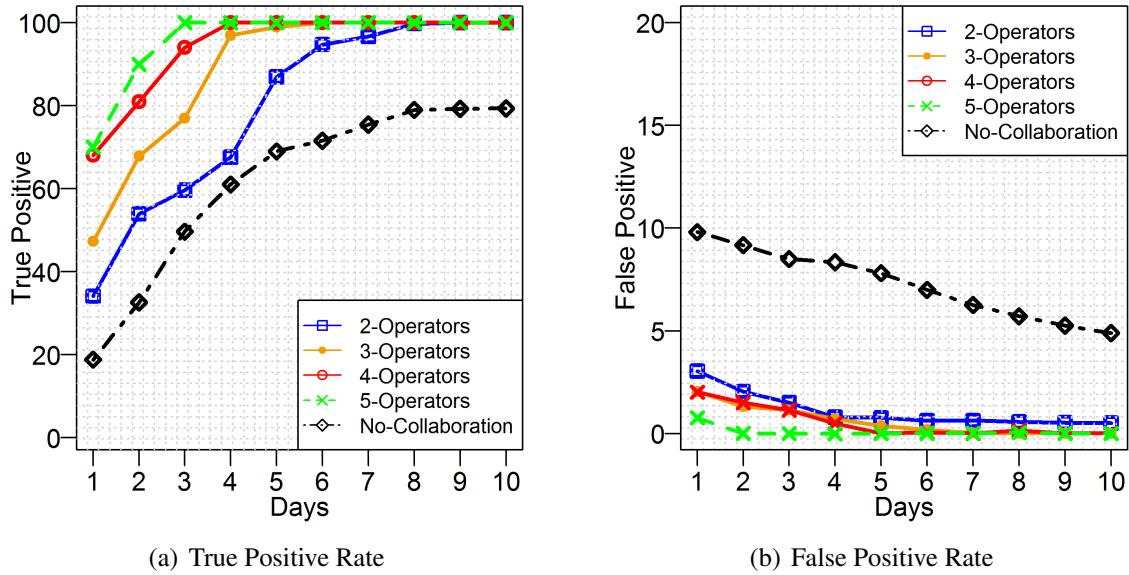


Figure 5.10: System Behavior Against Spammers Having High Out-Degree and High Duration Calls.

having spammers less than 10% as shown in Figure 5.9.B. The FP rate at low spamming rate can be improved further by using  $\beta$  value in combination with the number of unique callees of the subscriber. We recommend that  $\beta$  value should be between 2 and 3 and should be used in conjunction with the number of unique callees of the subscriber. The tradeoff between TP rate and FP rate for varied threshold is shown in a Figure 5.9.C. It can be seen from the Figure 5.9.C. that TP rate of COSDS increases with the increase in FP rate and is able to achieve 100% TP rate with relatively smaller FP rate.

### 5.6.6 Resilience Against Different Spam Calling Behaviors

It is possible that different spammers have different calling behaviors. In this simulation setup we consider three types of subscribers [CMP+13]: 1) subscribers calling a large number of callees, all their calls being successful and with a good duration, but without receiving any call from their callees. This would be a representative behavior of telemarketers and prank subscribers because of their large out-degree. 2) Subscribers calling a small number of callees per day, having only a few calls with good average duration, and also not receiving any call from the callees. These spammers always try to call a limited number of new callees within a specific time period. 3) Subscribers calling a small number of callees per day and manage to have high duration calls with many of them.

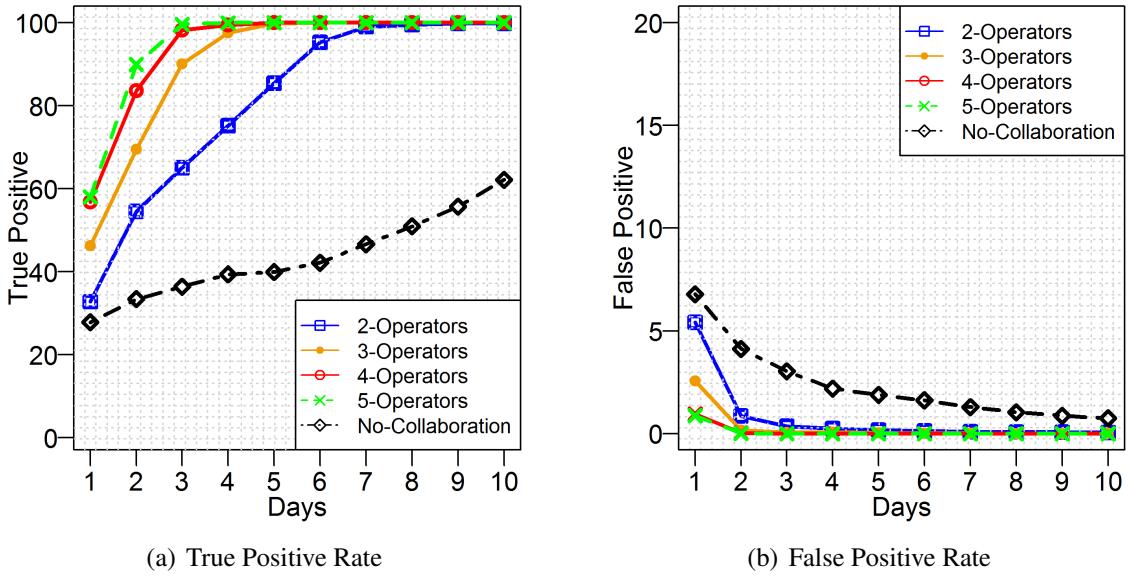


Figure 5.11: System Behavior Against Spammers Having Small Out-Degree and Small Duration Calls.

In the first experiment of this series we evaluated the system performance when spammers manage to have long duration calls to a large number of their callees (spam calling behavior 1). The experimental setup consists of 15000 spammers and 50000 legitimate subscribers equally distributed across five SPs. Each spam subscribers randomly choose a callee and calls to the 50% percent of the total number of customers in a SP. The call duration of these spam subscribers vary from 180 seconds to 400 seconds with an average call duration of 220 seconds. The legitimate subscribers in this experiment follow the same distribution as provided in section 5. The FP rates and TP rate of COSDS (collaboration with 3, 4 and 5 collaborators) is presented in Figures 8.A and 8.B. The FP rate decreases with the number of collaborators and at five collaborators COSDS achieves a FP rate of less than 0.5%. COSDS achieves TP rate of above 95% in six days much later than what it able to achieve with the spamming model presented in a section 5. This is because of the long duration calls of spammer to a large number of callees. Despite having high duration calls to a large number of callees, these subscribers are still identified as spammer because of their high number of callees and non-repetitive calling behavior. These subscribers also got small global reputation which also decreases over time.

In the second experiment of this series, we analyzed the resistance of the COSDS for the spam subscriber controlling his out-degree (spam calling behavior 2). In this experiment the spammer calls only 15-25 unique callees per day. The call duration varies from 90 seconds to 200 seconds with the average duration of 150 seconds. The TP and

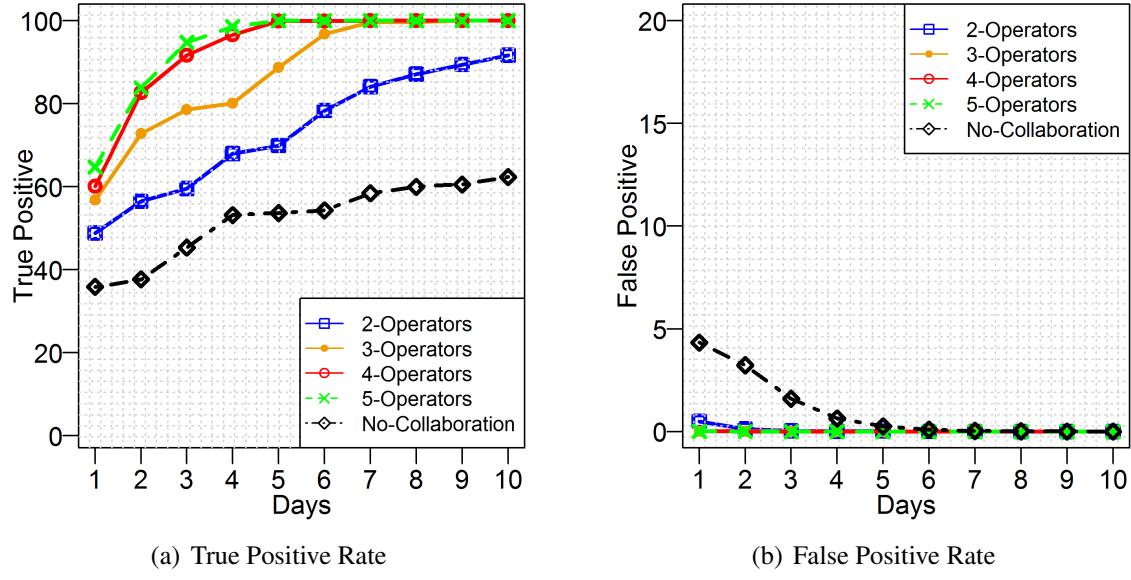


Figure 5.12: System Behavior against Spammers Having Small Out-Degree and Long Duration Calls.

FP rates of COSDS with 3, 4 and 5 collaborators is shown in Figures 5.11.A and 5.11.B. COSDS is able to block all spammers in 4 days, which shows that controlling out-degree would not allow spammer to call for the longer time period. As the spam subscriber increases the number of callees over time, his reputation scores starts decreasing as it does not receive many calls from his callees. The FP rate in this scenario is less than 0.5% from the day one.

In the third experiment of this series, we evaluated the performance of COSDS against spammers having large duration calls and small number of unique callees (spam calling behavior 3). In this scenario, the number of spammers and non-spammer are same as above. The spammers only calls 35-50 unique callees per day and the duration for each call is greater than 300 seconds, with average duration of 350 seconds. The TP and FP rate of COSDS with 3, 4 and 5 collaborating SPs are presented in Figure 5.12.A and 5.12.B. COSDS does not show effective resistance on first few days because of the fact that call duration of spammer is almost same as call duration of legitimate subscribers. However COSDS blocks almost all spammers after 6 days. The FP rate in this scenario is also decreasing and remains less than .5%.

From the Figures 5.10, 5.11 and 5.12, it can also be seen that non-collaborative system behaves poorly in terms of TP and FP rates for any type of spammers. From Figures 5.10, 5.11 and 5.12 we also conclude that spammers would be able to bypass COSDS system for relatively long time if they managed to have long duration calls without con-

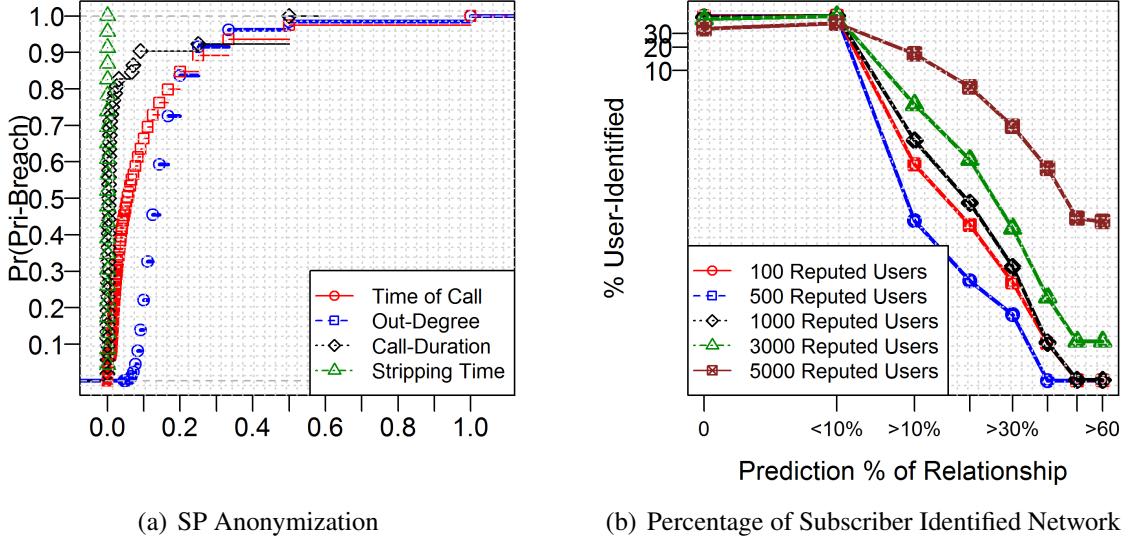


Figure 5.13: Privacy Breach Analysis for Different Scenarios: A) Probability of Breach for Some Auxiliary Information at SP; B) Percentage of Subscribers whose Relationship network identified to some percentage varying number of reputed subscriber.

trolling their number of callees or have to control out-degree. In order to remain undetected by COSDS, spammers need to control their number of callees with repetitive calling behavior.

### 5.6.7 Privacy Breach Analysis

One design consideration of COSDS approach is privacy preservation of relationship links of the subscribers. To estimate the possibility of privacy breach, we simulated the system for two privacy breach models: first, a privacy breach model where an adversary gets access to the local reputation engine and secondly, a privacy breach model where an adversary gets access to the global reputation scores of the subscribers at the CR.

In a first privacy breach model, the adversary learns some auxiliary information (call-rate, call-time, call duration, out-degree etc.) from the external source and intends to find the anonymized identity of the target subscriber so as to breach the relationship network of the target subscriber. Figure 5.13.A shows the CDF of privacy breach for the different auxiliary information that clearly shows that call time feature is more vulnerable to relationship privacy breach followed by the out-degree feature. We believe that collectively using the out-degree and call-time feature together would further increase the risk of a privacy breach. However, COSDS incorporates identity anonymization along with stripping of the call-time (stripping of seconds and minutes information from the call transactions)

thus reduces the risk of relationship privacy breach to almost zero using call-time as an auxiliary feature as shown in a Figure 5.13.A. However, it still vulnerable to the privacy breach using out-degree feature. The risk of privacy breach through use of out-degree auxiliary information can be minimized by using k-anonymized out-degree distribution but it will affect the detection accuracy.

In a second privacy breach model, an adversary obtains a part of highly reputed subscribers in a SP from the global reputation scores at CR and intends to make a guess for the relationship network of the target subscriber. The adversary creates predicted obituary relationship network of target subscriber by computing similarity between reputation score of target subscriber and other subscribers in a SP. If the predicted subscriber from obituary reputed list is also part of original relationship network of the target subscriber then we conclude that privacy breach has occurred to some extent. Figure 5.13.B shows the relationship level privacy breach percentages of COSDS approach as the number of subscriber in the adversary generated reputed list varies from 100 to 5000. The results from 5.13.B show the adversary would not be able to get at-least one friend of the target subscriber in his reputed list for more than 60% of the time. Normally, the percentage of relationship privacy breach increases with the increase in the number of subscribers in the obituary reputed list but it would not provide 100 % privacy breach even for high number of subscribers unless the adversary use all subscriber of particular SP in his reputed list. The number of friends in a reputed list can be too spare to form a detectable friendship network for the target. The adversary can guess with high probability that a part of relationship network of the target subscriber is a subset of his generated reputed list (say 10% has high probability shown in Figure 5.13.B), but he still does not have any clue which subscribers are common in adversary generated list and original target's friends list.

Another way to ensure the privacy of collaborators and their customers is to compute global reputation score through Secure Multi-party Computation (SMC) protocols. SMC enables collaborators to carry out computation task without revealing their reputation scores or private data. One such approach to perform encrypted computation is a homomorphic encryption scheme that performs computation on the encrypted data. There are a number of homomorphic schemes for example secure sum [CKV+02] and Paillier encryption [PAI99], which are homomorphic in summation and multiplication. However, the use of homomorphic approach would introduce overhead in computation process and further require evaluation for other performance metrics such as how privacy is protected in presence of malicious and honest but curious collaborators. Moreover, we believe that use of homomorphic encryption or secure sum would not affect the spam detection results of COSDS but would incorporate some additional overhead.

## 5.7 Discussion on COSDS System

The spammer may be able to bypass COSDS when it spams to only a few subscribers of the SP and then targets the SP again with a new identity. In this case the CR responds to SP with the same reputation score that the subscriber has within the SP. This can be overcome by binding the number of identities to the IP-address within the SP, linking many identities to same physical person or through imposing cost for a new identity. Linking identity to same physical person is a part of our future work. The spammer can also bypass COSDS by creating calling links between identities from multiple SPs. This allows spammers to call a reasonable large number of recipients but COSDS is able to block these spammers over the time. A new subscriber does not have social links and must be introduced within the SP in order to develop social relationships with their callees. If the SPs do not have any information about subscriber's local or global reputation then the SPs allow such subscriber to pass through network for few calls.

Although the percentile based detection shows resistance against top spammers, there is still much room for improvement with respect to collaboration with more features and classification approach for low- to mid-rank spammers. The detection approach can be improved by incorporating local social network features of the subscriber and a SPs-defined threshold along with the global reputation for final classification. This would minimize the FP rate caused by the low spamming rate and improve TP under high spamming percentage. COSDS can also implemented in a distributed way where the SP collaborate with their directly connected SPs in a privacy protection way. The COSDS approach can also be implemented with other reputation based approaches with slight changes and also involve callee for the final decision.

## 5.8 Conclusions

In this chapter, we have presented a collaborative SPIT detection system called COSDS, which employs collaboration among autonomous SPs for an accurate and early detection of spammers making low rate spam calls to recipients across many SP. The designed system requires collaboration with the exchange of non-sensitive summarized information to the trusted CR, which is not resource demanding regarding system and network resource, and is non-sensitive regarding subscriber's private information and operational aspects of SP. COSDS computes local reputation of the subscriber from subscriber's past call transactions within the SP and exchanges it to the CR for reputation aggregation and decisions about behavior subscriber. Our evaluation on synthetic data show that the COSDS

approach is very effective in detecting spammers, has reduced the detection time and provides privacy protection to the collaborating SPs. Specifically, it out-performs *SAS* in terms of detection time and accuracy, and achieves considerable same detection accuracy when collaboration is carried out through the exchange of CDRs and direct trust scores.

# Chapter 6

## EIS: Early Identification of Spammers

### 6.1 Introduction

In recent years, telecommunication networks have seen a dramatic increase in the number of subscribers around the world. According to GSMA statistics, there are more than 7 billion mobile users in total and more than 4 billion unique mobile users across the world. Malicious users can also acquire large number of identities to gain financial benefits because of spamming, advertisements and phishing attacks.

Users having multiple identities are able to evade the defense against spamming by pretending to be multiple, distinct individual in the network. Multiple identities can be misused to spread malicious information and spams. A spamming individual with multiple identities can use some of its identities to provide high recommendation to its other identities so to increase their reputation and evade detection system. Recent statistics revealed that Facebook has 1.23 billion active users and of which more than 90 million accounts are duplicate [FAC14]. This means that multiple accounts are owned by one individual and are mainly used for the malicious activities. A Phoneypot [GSB+15] study over seven weeks reveals that 36,912 unique phonetokens have received 1.3 million calls from the total of 252,621 unique sources. This can be attributed to the fact that telephony users receive large number of spam calls when scaled over a large number of users. Spammers in the telephony are more intrusive, require immediate response from call recipients and undermine the use of telephony for legitimate users.

Traditional reputation-based spam detection systems utilize user's identity for combatting spamming activities [KER11], [AM13], [TDZ+16]. If spammers make spam directly from one known identity to victims, they would be easily blocked by the spam detection system. However, spammers are adopting new ways to evade the spam detection systems

and stay undetected for unsolicited calls by acquiring large number of identities. Spammers can have a large set of calling identities for small cost and effectively obfuscates these calling identities to evade the spam detection systems. Although the spammer can have many identities but their victim network largely remains the same. Furthermore, spammer is not able to develop a strong social network with legitimate users from his different calling identities. The service provider requires an effective spam detection system against spammers having multiple identities in order to improve the experience of users, to ensure the trustworthiness of services provided and positively gain trust of customers.

Recently, several approaches have been proposed to link profiles of an individual across different and same social networks. These approaches estimate similarity between two profiles that belongs to one physical individual by using several features in three dimensions: 1) attempt to connect similar profiles owned by an individual by using social network structure of an individual across different social network platforms [ACF13], 2) attempt to estimate the similarity in profiles (age, sex, religion, location, name) of an individual on a two different social network platforms [VHS09, LWZ+14], and 3) attempt to measure the similarity in content posted by an individual on both platforms. In telephony, applying these features for connecting similar identities is not straight forward. Content-based similarity measure cannot be applied in telephony because content in telephony is speech which is resource intensive in terms of storing, retrieving and processing. As there is no profile information available in telephony, the feasible option left is social network connections of identities for estimating the similarity between identities. The information from user's social network connections have also been used for linking profiles of an individual from different social networks [ACF13], [JKJ13], but these approaches have only considered network connections and did not consider connectivity strength of an individual with others.

This chapter presents EIS (Early Identification of Spammers), a novel system that limits and blocks set of identities that are owned by a same spamming individual in a VoIP and voice networks. We exploit the notion that spammers can have multiple identities but they cannot establish strong connection with honest users and also have similar set of victims with similar calling behavior. The basic idea is that if the spammer has many identities, his social call graph become same in the sense that it has many common users between his identities. Our design is based on a use of call patterns and social network graph of identities. EIS system consists of three modules. 1) An ID-CONNECT module that connects similar identities that belongs to one physical person or individual by utilizing social network structure and calling behavior of identities. 2) A reputation computation module that computes reputation of an individual by using call-duration, call-rate and out-degree of the individual. 3) A spam detection module that flagged individual as

a spammer if reputation of the individual is less than automated dynamic threshold. To the best of our knowledge, our work is the first that utilizes weighted call graph for linking similar identities together in a voice network and uses identity linking for a spammer identification. We evaluate our proposed approach through experimental study using a synthetic graph dataset for the number of graph models and different spamming behaviors. We find that the proposed approach achieves high linking rate with a small candidate set size and is effective for the early identification of spammers frequently changing their identities. We believe that EIS system can be easily applied for connecting similar profiles in social networks by analyzing the social structure of profiles.

In summary, this chapter makes the following contributions:

- We introduce an ID-CONNECT system, a social network and behavior based model that links similar identities that probably belongs to a one physical individual. In ID-CONNECT, the weights on the links between identities are computed from the interaction rates and length of interactions. Individuals, especially spammers normally exhibit similar call behavior and have overlap in victims from their many identities. Two identities can only be considered as similar if they have common friends and have similar calling behavior towards common friends. ID-Connect is a two-step approach: firstly, it estimates the weighted similarity measure between identities by considering the call behavior of identities towards their common friends. Secondly, it generates candidate set for the given identity using fixed thresholds.
- A reputation engine that computes reputation of the individual by using call-rate, call duration and out-degree of the individual after connecting his identities. The input to the reputation engine is the data of linked identities formulated from the ID-CONNECT module. We believe that reputation computed after linking identities of an individual and analyzing his aggregate calling behavior would greatly separate spammers from the non-spammers.
- A detection module for computing automated classification threshold below which individuals are flagged as spammers. A dynamic automated threshold is being computed for each reputation cycle using the percentile based approach.
- We validate and evaluate EIS system through a comprehensive simulation study using a synthetic data set that we have generated using true behavior of spammers and non-spammers. The experimental results show that EIS system outperforms other identity linking systems and has shown effective resistance against spammers having many identities.

The rest of the chapter is structured as follows. Section 6.2 presents scenarios where users can have more than one calling identity and will provide definition of the identity linking and spam detection problem in a voice network. Section 6.3.1 overviews the procedure by which the proposed approach link identities together to be used for computing aggregate reputation of an individual and detection of the spammer. Section 6.4 explains our data generation process and presents some performance evaluation metric. Section 6.4 presents experimental results for different network models and different percentages of spammers. Section 6.5 provides discussion on an EIS system and finally, conclusions and future works are presented in Section 6.6.

## 6.2 Motivation and Problem Definition

In this section, we discuss the problem of spamming in a voice network, why legitimate users own more than one calling identity and define identity linking problem in a voice network. Furthermore, in section 6.2.5 we provide some necessary background definitions about social call graphs used towards design of identity linking system.

### 6.2.1 Spammer Network

In VoIP and other communication network (email, social network etc.), a legitimate user can receive spam calls and messages from several spam identities. It might be possible that several spamming identities are owned by one physical person or controlled by a one botnet. In email network, it is estimated that there exist more than 20% overlap in the distribution lists of the typical spammers [GCA+04] [JMG+09]. This large overlap is because of the fact that spammers normally use automatic ways to create the target identities, harvests identities or acquires identities of targets from some sources and launched spam attack from their different identities. The larger overlap in victim is also due to a fixed number of users in the network. In VoIP and mobile network acquiring a new identity is virtually not very costly, support plug and play, and can be easily integrated with the spamming systems over the Internet. A Phoneypot [GSB+15] study over two months reveals that 36,912 unique phonetokens have received 1.3 million calls from the total of 252,621 unique sources. This can be attributed to the fact that few unique sources share similar victims as many unique sources called phonetokens only once. The spamming model of the spammer is presented in a Figure 6.1.

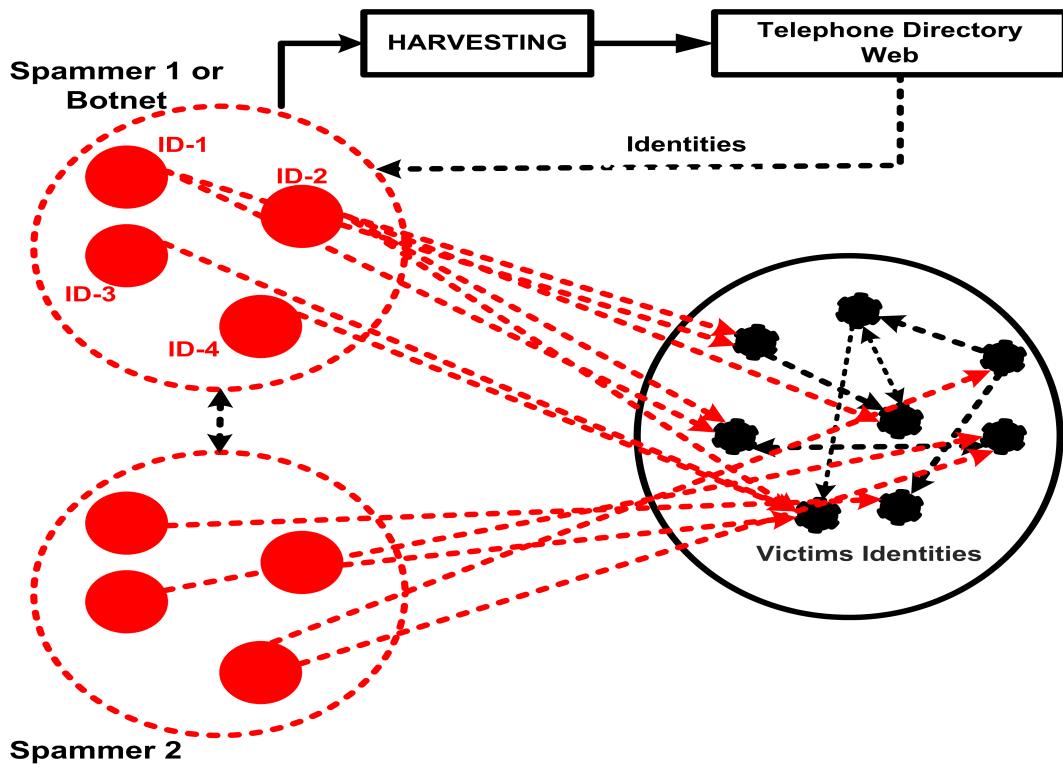


Figure 6.1: Attack Network of Spammer.

### 6.2.2 Why Users Have More Than One Identity

The number of active mobile connections around the world stands at more than 7 billion<sup>1</sup> and there are many countries where number of active cellular users are more than country's population. For example, Russia has 1.8 times more active cellular users than its population. This does not mean that every person in the country has exactly one mobile phone but can be attributed to a large number of people owning multiple calling identities. Legitimate users can have multiple calling identities for various reasons such as business and personal communications or to take advantage of cheap calling plans for the local and long distance calls. In addition to legitimate users, there are also non-legitimate users: users that are involved in illegally obtaining financial benefits by tricking people with scams. Spammers can purchase a large number of calling identities for free<sup>2</sup> or with minimal cost<sup>3</sup> and can use them from anywhere across the world with a simple Internet connection. Multiple calling identities can be misused to spread the malicious information and spams. The reputation of an individual is associated with the calling identity. A spamming individual can have many identities to defeat the reputation system by pre-

<sup>1</sup>GSMAIntelligence <https://gsmaintelligence.com/>

<sup>2</sup>google voice, inum

<sup>3</sup><http://www.voipfone.co.uk/>

tending that identities belong to multiple individuals. A spamming individual with many identities can also use some of its identities to provide high recommendation to its other identities so to increase their reputation and evade detection system. Furthermore the use of Do not call lists in many countries also induce system abusers to frequently change their identities so to reach their targets without being blocked.

### **6.2.3 Motivation**

A spammer can have more than one calling identity either by buying identities or faking identities of legitimate users. The spammer then uses these identities to target the legitimate users of the service provider. Connecting identities that belong to the same individual would provide detailed view about the behavior of the individual. In the spam detection domain, the linking of multiple identities of spammers would help in early identification of spammers and minimize the chances of Phishing and identity theft attack in a telecommunication network. One way to limit the number of identities to one individual is to cap the number of identities any single person can buy or impose some extra cost for buying identities. However, this still does not ensure that acquired identities would not be used for non-legitimate activities and it would also require a mechanism for linking similar identities if not properly linked at the time of identity purchase. However, in telephony identities can be acquired with minimal cost and plug and play is straightforward as users can buy SIM card or VoIP identity and start using. Another possible way to link identities is to link IP address in-case of VoIP and IMEI number (International mobile Equipment Identity) in-case of Mobile to link the identity and logs records of which identity belongs to which IP-address and IMEI. However acquiring new IP-address is not costly and the IMEI number can be manipulated. There is a strong need to have a system that automatically links similar identities together using calling behavior of identities and further use this to identify spammers and criminal rings.

The Spam detection systems presented in Chapters 4 and 5 decide about status of subscriber as a spammer or a non-spammer on the basis of observed call patterns for a specific subscriber identity. Since there is no identity linking, the spammer can easily bypass such system by simple changing their identity or whitewashing their global reputation system and rejoining the network. An important feature which can be used to distinguish spammers from non-spammers is the overlap in victims from spammer different identities and similar call behavior towards the overlapped victims. Beside spam detection, identity linking would also be effective for characterizing behavior of legitimate user for their different identities and identification of criminal groups.

### 6.2.4 Problem Definition

Every telephony service provider records call transactions of its customers in a Call Detail Record (CDR) database – a metadata that is used for a billing and network management. CDRs usually contain information including the calling identity of parties involved in a communication, the date and start time of calls, and the call duration of calls. In our problem setting, raw CDRs are used to build a directed graph  $G = (V, E)$  to represent a call network of identities  $V$ . In addition to network structure, we also have weights on the edges  $E$  derived from the call statistics between identities such as the call rate and the call duration. For example, in a voice social call graph, nodes are the identities of individuals registered in a network, edges represent who calls who, and weights on edges represent connection strength between nodes. An individual can have more than one calling identity with a separate call graph for each calling identity. Let  $P$  denote the set of all individuals registered in some service provider ( $SP$ ) and there exists some individuals those have more than one calling identity. Consider the set of all identities that belong to a single individual  $\{P_i\} (P_i \in P)$ . We define identity linking problem as follows.

**Definition 1 (Identity linking Problem):** Let  $GT_1$  and  $GT_2$  be the social call graphs of all identities at time periods  $T_1$  and  $T_2$ . The identity linking problem requires that given some identity from  $GT_2$ , to find a set of identities from  $GT_1$  that are similar to the given identity. When  $Sim(ID(GT_2), ID(GT_1)) > threshold$  then identities are considered as similar to each other.

The process of a identity linking is depicted in a Figure 6.2 and can be defined as follows:

$$IL(ID_2, ID_1) = \begin{cases} True & \text{if } ID_2 \text{ and } ID_1 \text{ belong to } P_i \\ False & \text{otherwise} \end{cases}$$

The identities from  $GT_1$  can be added to the candidate set of the given identity if the similarity score is greater than a defined threshold.

Once similar identities that belong to one physical individual are linked, the next step is to compute the aggregate reputation of an individual by considering call statistics from all his identities and then further makes decision whether individual is a spamming or not.

**Definition 2 (Spam Decision Problem):** Let  $P$  denote the set of all physical individuals registered in a service provider ( $SP$ ). There are some individuals that have more than one calling identity. Consider the set of all identities that belong to the single individual  $\{P_i\} (P_i \in P)$ . We define spam detection problem as follows.

$$Decision(P_i) = \begin{cases} Spam; \text{if } Reputation(P_i) < Threshold \\ Non-Spam; \text{if } Reputation(P_i) > Threshold \end{cases}$$

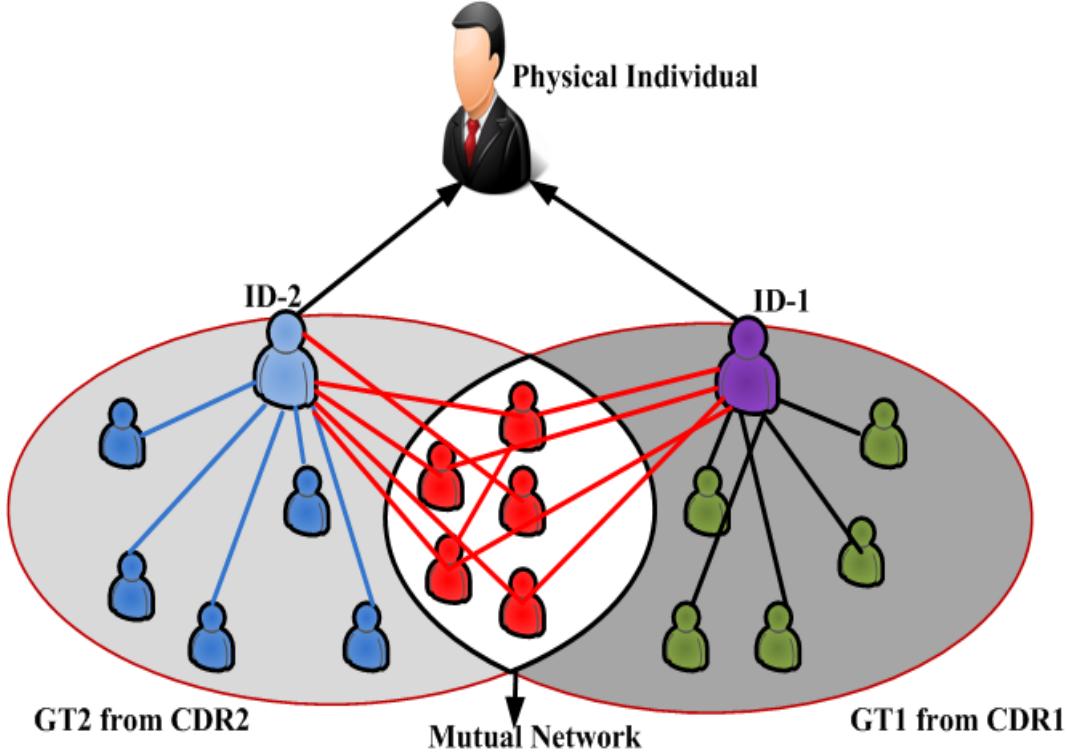


Figure 6.2: The Calling network of physical individual for two different time periods with two different identities  $ID_1$  and  $ID_2$ .

### 6.2.5 Background Definitions

Telecommunication operators record transactions of their users in a call detail record (CDR) database, which is normally available in the form of plain text. In order to formalize our social network based identity linking system, a call graph is to be created for each unique identity in the raw CDRs. In this section we provide definitions of terms and notations used for the design of identity linking system.

**Definition 3 (Identity):** A calling identity is the telephone number or VoIP identifier that identifies a user in a service provider and is particularly used to make and receive calls. The service provider records every transaction of a user under this identity. An identity can be a caller  $S$ , a callee  $R$ , or both.

**Definition 4 (Call Detail Records):** A CDR represents a call transaction record between a subscribers and can be mainly defined with five major fields:  $(S,R,T,Du)$ ; where  $S$  and  $R$  are the identities involved in a call,  $T$  is time when the call initiated, and  $Du$  is the duration of a call.

**Definition 5 (Call Graph):** A call graph  $G(V,E)$  is a directed graph where  $V$  is an identity and  $E$  represents edges between identities. If the number of identities in  $V$  is  $n$ , then the adjacency matrix graph is an  $n \times n$  matrix in which entry  $A_{SR}$  is equal to 1 if  $(S,R) \in E$ , and 0 otherwise.

**Definition 6 (Friendship Network):** Let  $R_{in}$  be the set of identities from whom  $R$  receives calls and  $R_{out}$  be the set of identities to whom  $R$  makes calls. The combined network is the reunion of the sets of in and out friends of an identity is  $(R_{in} \cup R_{out})$  and can be represented as  $R \subseteq P$ .

**Definition 7 (Weighted Call Graph):** A weighted call graph is a call graph where the edges between identities are assigned non-binary weights. The weighted graph can also be represented as an adjacency matrix  $WG$  where  $WG_{SR}[0,1]$  is the weight of link between identity  $S$  and identity  $R$ . In a voice network these weights are assigned from the call frequency, call duration, or both. In this chapter we use definition 8 below.

**Definition 8 (Average Call Duration  $W_{SR}$ ):** In ID-CONNECT, the edge weights  $W_{SR}$  are computed from the average call duration. The average call duration between  $S$  and  $R$  is computed by dividing the sum of the duration of all calls ( $CD_{SR}$ ) by the total number of calls made between  $S$  and  $R$  ( $CallRate_{SR}$ ) during a given period.

$$W_{SR} = \frac{\sum CD_{SR}}{CallRate_{SR}} \quad (6.1)$$

**Definition 9 (Mutual friends):** Let  $GT_1$  be the friendship network of identity  $ID_1$  at  $T_1$  and  $GT_2$  be the friendship network of identity  $ID_2$  at  $T_2$ . The mutual friends  $MF(ID_2, ID_1)$  between  $ID_2$  and  $ID_1$  can be computed as follows:

$$MF(ID_2, ID_1) = F_{ID_2} \cap F_{ID_1} \quad (6.2)$$

where  $F_{ID_1}$  and  $F_{ID_2}$  denote the set of friends of  $ID_1$  in  $GT_1$  and  $ID_2$  in  $GT_2$ , respectively.

## 6.3 EIS: Early Identification of Spammers through Identity linking and Reputation Aggregation

In this section, we present an EIS system. The goal of an EIS system is to link identities that belongs to a physical individual; computes reputation of an individual, and classifies an individual as a spammer and a non-spammer. Specifically, EIS system consists of three modules: an ID-CONNECT module for linking identities that belongs to a physical individual, a reputation module which computes reputation of an individual and a classification module which computes automated threshold below which an individual is classified as a spammer. The system architecture of an EIS system is shown in a Figure 6.3 and modules are detailed as following.

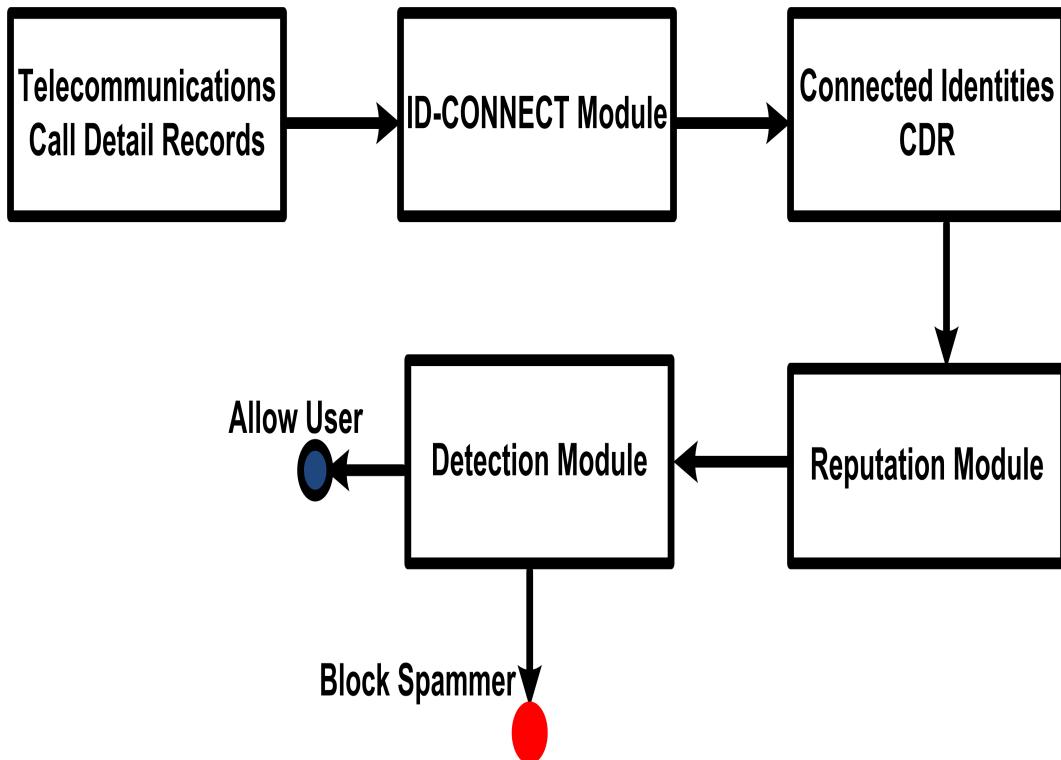


Figure 6.3: Building Block of EIS System Consisting of Three Major Modules.

### 6.3.1 ID-CONNECT Module

Given CDRs for the two time periods, the goal of an ID-CONNECT module is to identify similar identities that probably belong to a physical individual. To this extent, ID-CONNECT module uses social and call features of identities to link similar identities. The building block of an ID-CONNECT module is shown in a Figure 6.4 and procedures for linking identities are presented in an algorithm 6.1 which are called sequentially. The rest of this section explains the following three concepts.

**Construction of a Call Graph:** The call graph of identities is constructed from the raw CDRs of two time periods. Section 6.3.1.1 provides details of the process used for the construction of the call graph.

**Weighted Similarity Measure:** The weighted similarity between a given identity from time  $T_2$  and all identities from time  $T_1$  is computed by considering calling behavior and social connections of identities towards friends. Section 6.3.1.2 provides details on the procedure used for estimating similarity between identities.

**Candidate Set:** Once a weighted similarity between identities has been estimated, the next step is to generate candidate set for the given identity from  $T_2$ . Section 6.3.1.3 provides details on the process used for finding Candidate List(CL) of the given identity.

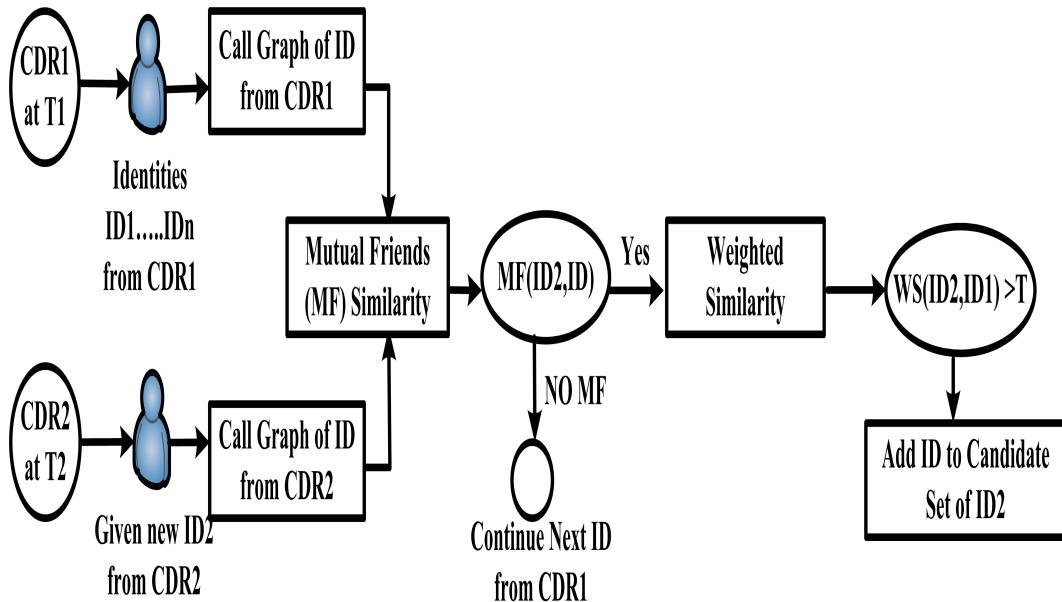


Figure 6.4: The work-flow of ID-CONNECT for identity linking. The approach outputs list of all identities from  $CDR_1$  which are similar to a given identity from  $CDR_2$ .

### 6.3.1.1 Call-Graph Construction

A CDR contains all the information about the call transaction such as time of call, duration of call, who originated the call, who received the call etc. The CDRs are logged at the call processing engine or the proxy server and stored in the CDR data server. In this chapter we consider CDRs of identities for two time periods  $T_1$  and  $T_2$  and construct a weighted social call graph. From the processed CDRs if we consider node in a social call graph as an identity of the subscribers, and a call transaction between identities as the link connecting the two nodes. We can construct a weighted social call graph  $G = (V, E, W)$ , where  $V$  is the set of nodes associated to each identity, while  $E$  is the set of link between nodes, and  $W$  is the weight on the link between nodes and is computed from the call duration and call rate between nodes. In this chapter a non-negative weight  $W_{SR}$  is computed from the average call duration. After constructing the weighted call graph our goal is to estimate the similarity between identities and generate candidate set for the given identity.

---

**Algorithm 6.1** Extraction of Candidate-Set

---

```

1: procedure EXTRACTION OF MUTUAL FRIENDS ()
2:   input  $\leftarrow$  Call Detailed Records at Time T1 and T2.
3:   output  $\leftarrow$  Mutual Friend List.
4:   Extract New Identities that appears at T2.
5:   Extract List of Identities from T1 which have mutual friends to a given identity
   from T2.
6:   for each Identity from T2 do
7:     if  $MFS(ID_2, ID_1) > 0$  then
8:       add Identity to MFL(Mutual friend list).
9:     else
10:      do not add Identity to MFL.
11:    end if
12:   end for
13: end procedure
14: procedure ESTIMATING WEIGHTED SIMILARITY ()
15:   input  $\leftarrow$  MFL.
16:   output  $\leftarrow$  SIMList of Identities for a Given Identity.
17:   for each User Identity and identities in MFL do
18:     Compute Weighted Similarity between given Identity and identities in a MFL
   using their mutual friends and equation 6.4.
19:   end for
20: end procedure
21: procedure FINAL CANDIDATE-SET()
22:   input  $\leftarrow$  SIMList.
23:   output  $\leftarrow$  Final Candidate-Set for a given identity.
24:   for each Identity and his SIMList do
25:     if  $WS(ID_{T2}, ID_{SimList}) > threshold$  then
26:       add Identity to the Candidate-Set().
27:     else
28:       do not add Identity to the Candidate-Set().
29:     end if
30:   end for
31: end procedure

```

---

**6.3.1.2 Estimating Similarity**

A common way to estimate similarity between identities is to compute the number of mutual friends between identities. The identities are considered similar if they have a large number of mutual friends. Given a  $GT_1$  a friendship network of identity  $ID_1$  at  $T_1$  and  $GT_2$  a friendship network of identity  $ID_2$  at  $T_2$ . The mutual friend similarity  $MFS(ID_2, ID_1)$  between  $ID_1$  and  $ID_2$  can be computed by normalizing the size of  $MF(ID_2, ID_1)$  (as

computed in equation 6.2) to the size of the union of the set of friends of  $ID_1$  and of  $ID_2$ .

$$MFS(ID_2, ID_1) = \frac{|MF(ID_2, ID_1)|}{|F_{ID_2} \cup F_{ID_1}|} \quad (6.3)$$

The similarity measure based on the mutual friends does not consider the strength of the relationship between identities and their friends. The number of mutual friends between identities increases with the size of friendship network and consequently a larger candidate set is returned for a given identity. This would result in a high false positive on a small similarity threshold. In general, similar identities not only have common friends but also show similar call behavior towards them.

The calling behavior of identities towards mutual friends can provide additional information about which two identities are more similar to each other. For instance, consider an example of a call graph presented in Figure 6.5, where our goal is to find candidates for an identity  $ID_2$ . Identity  $ID_2$  has 6 friends, whereas identity  $ID_1$  and  $ID_3$  has three friends each, all of them are common to the friends of  $ID_2$ . The mutual network similarity of  $ID_2$  to  $ID_1$  using equation 6.3 is thus  $MFS(ID_2, ID_1) = 0.5$ , and network similarity of  $ID_2$  to  $ID_3$  is  $MFS(ID_2, ID_3) = 0.5$ . In MFS, both  $ID_1$  and  $ID_3$  are added to the candidate set for the identity  $ID_2$ . However, we believe that identities should be considered as candidate of a given identity not only if they have mutual friends but also if they have similar call behavior towards mutual friends. The size of candidate set can be reduced by computing weighted similarity measure between identities.

**Definition 10 (Weighted Similarity Measure):** The weighted similarity  $WS(ID_2, ID_1)$  is computed as follows.

$$WS(ID_2, ID_1) = \frac{1}{OD(ID_2)} \sum_{k \in MF} \frac{\min(W_{ID_2^k}, W_{ID_1^k})}{\max(W_{ID_2^k}, W_{ID_1^k})} \quad (6.4)$$

In equation 6.4,  $W_{ID_2^k}$  is the edge weight of  $ID_2$  to his  $k^{th}$  friend,  $W_{ID_1^k}$  is the edge weight of  $ID_1$  to his  $k^{th}$  friend,  $OD(ID_2)$  is the out-degree of  $ID_2$  and  $k$  is the number of mutual friends between  $ID_1$  and  $ID_2$ .  $WS(ID_2, ID_1)$  is between 0 and 1. The more similar the identities are in terms of common friends and behavior, the closer the weighted similarity is to 1. On the contrary, the more dissimilar the behavior of identities towards mutual friends, the closer the value of the weighted similarity will be to 0. Consider again the example of the call network presented in Figure 6.5 with edge weights. The weighted network similarity between  $ID_2$  and  $ID_1$  is greater than the similarity score of  $ID_2$  and  $ID_3$ . This is because edge weights of  $ID_2$  and  $ID_1$  towards their mutual friends are extremely similar, whereas edge weights of  $ID_2$  and  $ID_3$  are quite different.

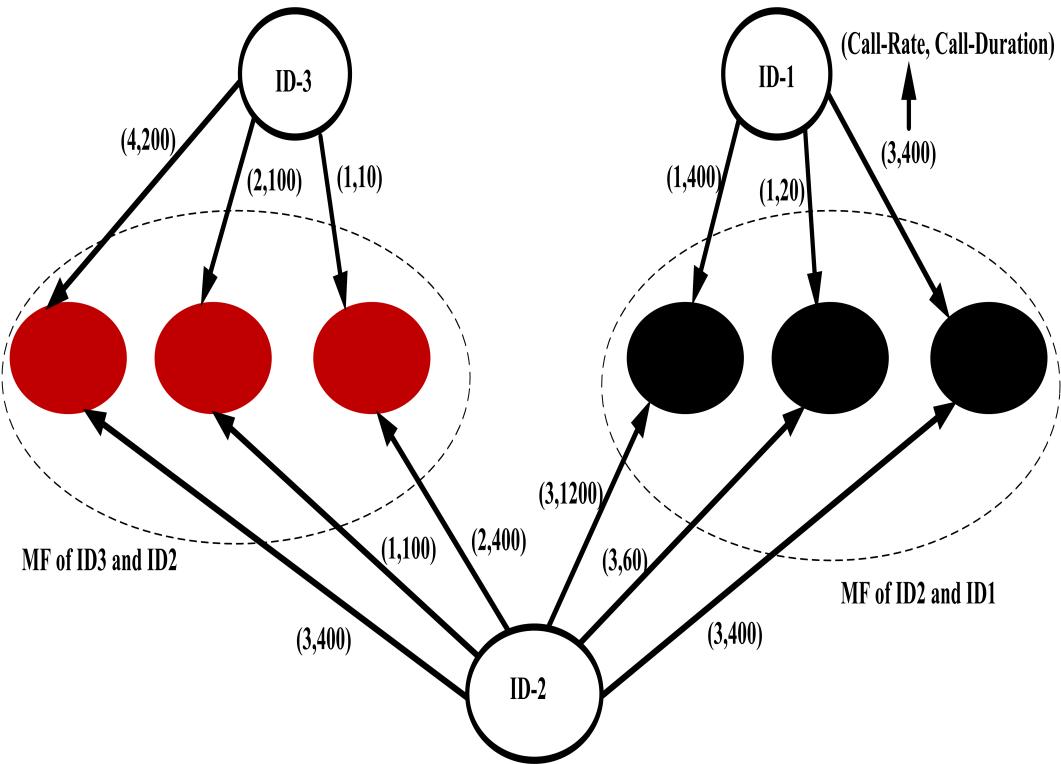


Figure 6.5: An example Weighted Call Graph for a given identity  $ID_2$  from  $T_2$  and two identities ( $ID_1$  and  $ID_3$ ) from  $T_1$ . Without link weights  $ID_2$  is similar to  $ID_1$  and  $ID_3$  but when link weights are considered then  $ID_2$  is more similar to  $ID_1$  than  $ID_3$  because of similar link weights. For ease of reading, rather than showing the  $WG_{SR}$  weight directly on each edge, we show vector  $(CallRate_{SR}, \sum CD_{SR})$ .

Specifically, using equation 6.4,  $WS(ID_2, ID_1) = 0.5$ , whereas  $WS(ID_2, ID_3) = 0.158$ . With the weighted similarity measure, only  $ID_1$  is considered a candidate for an identity  $ID_2$ . If two identities have no mutual friends their similarity is zero.

### 6.3.1.3 The Identity-Linking Process

The identity linking process decides whether an identity in  $T_1$  should be included in the candidate list of a given identity in  $T_2$  is as follows.

**Definition 11 (Similarity Threshold):** Each identity from  $CDR_1$  should be added to the candidate list of a given identity from  $CDR_2$  if weighted similarity measure between identities is greater than a threshold  $T$ .

$$CL(ID_2) = \begin{cases} ID_1 \in CL(ID_2) & \text{if } WS(ID_2, ID_1) > T \\ \emptyset & \text{if } WS(ID_2, ID_1) < T \end{cases}$$

A large similarity threshold would result in a small number of identities in  $CL$  because of the requirement of maximum network and behavior similarity. This however, would

also likely not include the identity that correctly links to the given identity. A small threshold would likely include the identity that correctly links to the given identity, but would include a large number of identities in the candidate set. The threshold needs to be chosen by having trade-off between true positive (correct linking) and false positive (incorrect linking). The candidate set can also be processed by a speech processing engine for verification and exact matching, but at high cost. The ideal solution is that the identity linking process produces a candidate set for the given identity and then speech processing engine would perform speech matching process on a small candidate set, thus minimizing the time required for the exact match.

### 6.3.2 Reputation Module

Reputation management is necessary to assess the trustworthy behavior of an individual for blocking the spammers from calling, so to minimize the interaction of user with the malicious and spamming users. The reputation of an individual is computed by aggregating the direct trust scores between individuals. Trust between individuals can be estimated by simply considering feedback from the individuals about their call interactions or can be estimated from call records in an automated way. In former approach, the individual rates other individual after the end of their call transaction, however this approach require changes in phone set and is also intrusive, thus is not feasible in actual deployment. The later approach computes trust using information from the call detailed record. Though these systems are non-intrusive, but are considering one feature for the trust computation which can be easily evaded by spammers [AM13]. The reputation of an individual is then computed by aggregating the direct trust scores (feedback or trust from CDRs) .

The reputation of an individual is computed using identity of the individual. An individual can have one or more identity. If an individual have many identities and make calls separately this would affect the reputation system as it provides incorrect view about an individual. The solution to this issue is first link identities of an individual and then computes reputation of an individual considering all his identities. In previous section we presented approach for identity linking, in remaining of this section we outline an approach used for computing trust and reputation of an individual.

In a voice network of  $n$  individuals with the linked calling identities, the service provider computes trustworthiness of individual behavior towards others by the direct trust scores between individuals from their call transactions. Consider a trust matrix  $Trust_{SR}$ , where  $T_{SR}$  is the trust score between an individual  $S$  and the individual  $R$ . If there is no interaction from  $S$  to  $R$  then  $T_{SR}$  is set to 0. The trust between an individual  $S$  and an individual  $R$  is computed by using call statistics extracted from the past call

transactions between  $S$  and  $R$ . We use the following features for computing direct trust between  $S$  and  $R$ : the incoming and out-going frequency and call duration between  $S$  and  $R$ , and the out-degree of  $S$ . Specifically, given a call duration matrix, the call-rate matrix and out-degree vector of all individuals, the direct trust between  $S$  and  $R$  is estimated using equation 6.5.

$$Trust_{SR} = \frac{CD_{SR} \times CallRate_{SR} + CD_{RS} \times CallRate_{RS}}{PO_S} \quad (6.5)$$

The trust matrix  $Trust_{SR}$  is generally sparse as individual usually interacts with only few other individuals. In equation 6.5  $CD$  is the  $n \times n$  call duration matrix of  $n$  individuals,  $CallRate$  is the  $n \times n$  interaction matrix of  $n$  individuals and  $PO$  is the out-degree of  $n$  individuals. The direct trust computed from the equation 6.5 would result in a small direct trust scores for the individuals who have large number of short duration calls, have large number of friends, and only receive few small duration calls from interacted individuals. On the other hand, individuals who have long duration repetitive calls would manage to maintain good trust towards their called individuals. From equation 6.5 we can infer that the legitimate individuals would achieve high trust score with a large number of their interacted individuals and spammer would have a small trust scores with a large number of other individuals.

---

### Algorithm 6.2 Reputation Computation

---

```

1: procedure GLOBAL REPUTATION OF ALL USERS  $\{S\}$ 
2:   input  $\leftarrow Trust$  (normalized direct trust matrix, with elements  $T_{SR}$ )
3:   output  $\leftarrow GR$  (Global reputation score vector, with elements  $GR_S$ )
4:   precision parameter  $\leftarrow \varepsilon$ 
5:   Initialize reputation vector  $GR$ 
6:    $GR_S = [1/PO_S]$ 
7:   Iterate until convergence
8:   while  $\delta < \varepsilon$  do
9:      $GR \leftarrow Trust \times GR$ 
10:     $GR \leftarrow GR / \|GR\|$ 
11:     $gr \leftarrow \|GR\|$ 
12:     $\delta \leftarrow \frac{gr - gr_{previous}}{gr}$ 
13:     $gr_{previous} \leftarrow gr$ 
14:   end while
15: end procedure

```

---

Once the direct trust between individuals has been computed, the next step is to compute the global reputation of the individual by aggregating the direct trust score of an individual towards all his interacted individuals. For the global reputation aggregation,

the direct trust score of an individual must be normalized which is defined as:

$$T_{SR} = \frac{Trust_{SR}}{\sum_R Trust_{SR}} \quad (6.6)$$

The normalized  $T_{SR}$  has values between 0 and 1, where each row sums to  $\sum_R T_{SR} = 1$ .

Consider a normalized direct trust matrix  $Trust$  from equation 6.6. Let  $GR$  be the global reputation vector of an individual  $S$  which is computed from the iterative method specified in algorithm 1 and presented below. The global reputation scores of all individuals are then computed by performing the following iterative operation on the normalized direct trust matrix and initial reputation vector.

$$GR(t+1) = Trust \times GR(t) \quad (6.7)$$

Initial the global reputation score is set to  $1/PO_S$  i.e. if individuals have huge out-degree they should assign small initial reputation scores. The iterative process of an equation 6.7 continues until norm of  $GR$  that is  $\delta = \frac{gr - gr_{previous}}{gr}$  is less than the predefined threshold  $\varepsilon$ .

### 6.3.3 Spam Detection Module

The reputation of a SPIT individual deviates much from the reputation of legitimate users and could be used for distinguishing SPIT from the non-SPIT. In this subsection, we show how SPIT individual can be distinguished from non-SPIT using global reputation scores and an automatically computed threshold. A percentile based automated threshold is computed using following method. First, the reputation scores of all users are sorted in descending order and the threshold value is set to the 25th percentile of set. Second, the mean of the individuals with global reputation less than the 25th percentile value is set as a dynamic threshold for a specific time window. The procedure for classifying individuals as SPIT or non-SPIT is presented in an algorithm 2 of Chapter 4. Individuals can be classified as legitimate 1 or non-legitimate -1 based on a following rule:

$$Individual_P = \begin{cases} GR_P > \beta \times \text{threshold} & ; 1 \\ GR_P < \beta \times \text{threshold} & ; -1 \end{cases}$$

## 6.4 Experimental Data Set and Evaluation Parameters

The EIS methodology is experimental evaluated in this section. First, we verify if ID-CONNECT module can accurately link the similar identities together, comparing with

some other baseline identity linking solutions. Second, we examine the effect of identity linking towards early identification of spammers.

### 6.4.1 Analysis of ID-CONNECT

In this section we validate ID-CONNECT module using a random data set generated using different network models and different percentages of overlapped network. First, we detail the process that has been followed for generating synthetic data set and identify the evaluation parameters.

#### 6.4.1.1 Identity Linking Data Set

There are several ways for producing synthetic graphs that can be used for generating a call network structure similar to that of a real voice operator. In order to address all possible network structures, we generated data for the three well known random graph models. 1) The Erdős Rényi model (a.k.a. ER model) [NEW05b] generates a random graph in which edges between nodes have equal probability of existence. ER random graph is different from real-world networks in two aspects i.e power law degree distribution and clustering coefficient. The power law degree distribution has been adopted in a Barabási-Albert model(a.k.a. BA model) [NEW05b], which considers the phenomena of growth and preferential attachment, yet clustering coefficient is still not adequate in a BA model. The Watts-Strogatz (a.k.a. WS model) [NEW10] considers clustering coefficient, power law degree distribution, shortest average path length, and also exhibits small-world effect. Many of the voice telecommunication operators have power law degree distributions (with  $2 < \alpha < 3$  power law parameter), which means that a large number of nodes have small node degree whereas a small number of nodes have very high node degree [NEW05a], [SMS+08].

We generate random networks for 10000 users with fixed parameters for edge probability and minimum number of friends. In the ER model the edge probability between nodes is 0.2 and the minimum number friends is greater than 10. In the BA and WS models the average number of friends for a given node is greater than 10 with the probability of connection between nodes is 0.5. Once a call network of each node has been generated, the next step is to assign weights on edges between nodes. Telephony users typically show Poisson distribution [CHA2015] for their call rate; this has been used in our simulations with average call rate of  $\mu = 3$  calls (cf. equation 6.8).

$$CallRate_{SR} = \frac{e^\mu \mu^x}{x!} \quad (6.8)$$

Call Duration characterizes the length of communication between subscribers. Users normally have long duration calls with friends and short duration calls with their colleagues or non-friend, but have exponential distribution on aggregate [YL07] [DDT+09]. We use exponential distribution for the call duration with  $\mu = 360$  seconds (cf. equation 6.9).

$$p(\text{CallDuration}_{SR} = x) = \mu e^{-\mu x} \quad (6.9)$$

Finally, the data is generated for two time periods. During first time period, CDRs for 10000 users is generated. For the second time period, we generate data by changing the identity of 30% of users while the remaining 70% also appear in second time period. The call-rate and call duration distribution for users of both time periods is same. A new identity of the same user in  $T_2$  has some overlap in the set of mutual friends from  $T_1$  (from 1 mutual friend to complete overlap, uniformly randomly chosen) and has the same call rate and call duration distribution parameters of his previous identity in  $T_1$ . In order to report average results we repeated each experiment 10 times.

#### 6.4.1.2 Evaluation Parameters

We analyzed effectiveness of ID-CONNECT module using two parameters: the True Positive Rate and the Candidate set size.

**True Positive Rate (TPR):** TPR is the percentage of the number of times a given identity is correctly linked with the identity from the other time period.

**Candidate Set Size (CSS):** CSS is the number of candidate identities that are returned for the given identity. We add identities to a candidate set of a given identity if the similarity score of an identity from  $T_1$  and given identity from  $T_2$  is greater than a threshold.

#### 6.4.1.3 Other Approaches from the Literature

In next section, we compare the ID-CONNECT (labeled as IDC in evaluation graphs) results with the following similarity measure approaches.

**Out Mutual Friends:** The Out Mutual Friends is number of common out friends linked to the focused identities and is a most basic neighbor-based approach for similarity estimation. Various approaches, such as Jaccard similarity, cosine similarities etc. have been used for estimating mutual friend's similarity between nodes in a graph. In this chapter, we use Jaccard similarity measure for similarity estimation between identities and is labeled as OMF in graphs presented in the next section.

**Out Mutual Friend’s Connections:** The Out Mutual Friends Connections [ACF13] considers the friendship links among the mutual friends of the focused identities in addition to the number of mutual friends between focused identities. In next section, Out Mutual Friends Connection is labeled as CON.

**Out-SimRank:** SimRank [JW02] considers the entire graph structure to determine the similarity between focused identities rather than by considering the neighbors of the focused identities. In analysis, we considered out-friends of focused identities for recursively computing similarity between them. We used following equation for Out-SimRank similarity between the focused identities i.e ID2 and ID1.

$$S(ID2, ID1) = \frac{C}{DM} \sum_{i=1}^{|O(ID2)|} \sum_{k=1}^{|O(ID1)|} S(O_i(ID2), O_k(ID1)) \quad (6.10)$$

In equation 6.10,  $DM = |O(ID2)||O(ID1)|$ ,  $O$  is the out-degree of identities,  $C$  is the constant between  $[0,1]$  and  $(O_i, O_k)$  is the out friends of friends of ID2 and ID1. In next section, graphs for Out-SimRank are labeled as OSR.

We believe OMF and CON are the competitors to the IDC because both use direct neighbors of the focused identities for computing similarity between focused identities. The approaches based on mutual friends are not computational complex, however they may have different true positives and candidate set size depending on the inclusion or no inclusion of edge weights. On the other hand, OSR compares focused identities mutual friends along with the entire graph structure and is computational complex compared to approaches based on common friends.

#### 6.4.1.4 Performance Results

In this section, we analyze the performance of ID-CONNECT for different similarity thresholds and compare it with above mentioned approaches.

##### 1) Threshold Variation, BA model :

We analyze the performance of the different approaches for the similarity threshold varied from 0.1 to 0.9. A threshold of 0 means that the ID-CONNECT system is required to consider every identity similar to the given identity, and thus the candidate set includes all identities. Such a small similarity threshold would include many wrong identities. A threshold of 1 means that a complete overlap in friendship network and a similar behavior is required for considering two identities as similar. Such a high similarity threshold would likely not include any identity in the candidate set. We highlight evaluation results in a bold for the candidate set greater than 3 and TPR greater than 80%. Figure 6.6

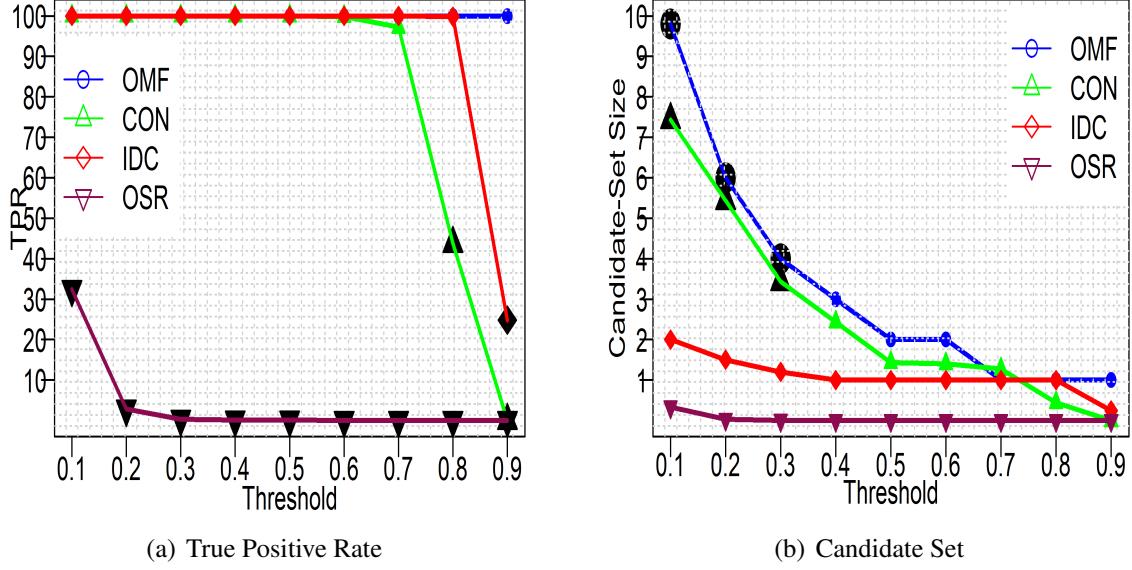


Figure 6.6: Performance Results for Different Threshold and Barabási-Albert.

demonstrates that use of behavior patterns and weighted network for the similarity estimation would greatly decrease the candidate-set size without affecting the TPR. From Figure 6.6.A it is clear that TPR is almost same for all approaches for any similarity threshold. As threshold increases, the TPR of ID-CONNECT decreases to 20% for a threshold of 0.9, as shown in Figure 6.6.A. The decrease in TPR of ID-CONNECT is due to the fact that a difference in a call rate and call duration between identities towards their mutual friends would result in a small similarity scores between identities despite having mutual friends. Considering weights for similarity measure also minimizes the size of candidate set at the small thresholds without effecting TPR.

The TPR of all approaches behaves similarly up to threshold of 0.8. However, TPR should be analyzed in connection with the size of the candidate set. The ideal approach should have high TPR with a small candidate set. It is obvious that candidate set size decreases with the increase in threshold as shown in Figure 6.6.B. The large number of identities in a candidate set consumes more time and requires more resources in searching for the correct match of the given identity with techniques such as speech processing. From Figure 6.6.B., we observe that for most threshold values ID-CONNECT returns only one identity for the given identity; this is the correct identity. On the other hand, OMF has 5-9 identities in the candidate set when the threshold is smaller than 0.5 and has candidate set size of less than 3 when the threshold approaches 1. This means that OMF only returns the correct identity along with a small candidate set when the threshold is high and approaching 1.

The small candidate set of ID-CONNECT is due to fact that it only considers those

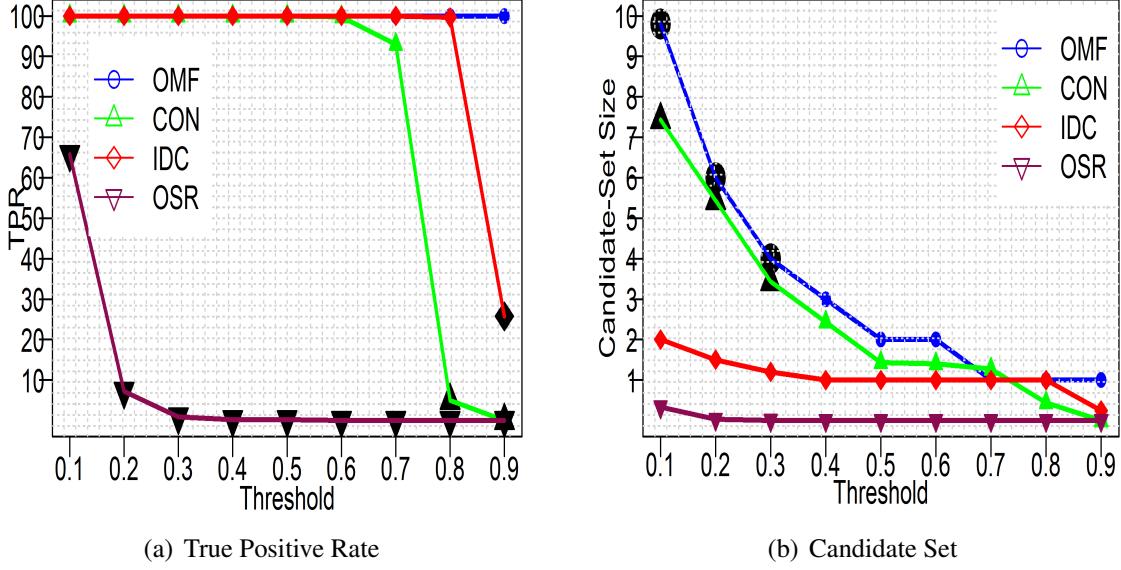


Figure 6.7: Performance Results for Erdös Réenyi.

identities as similar if identities have similar call behavior towards their common friends. Analyzing Figures 6.6.A and 6.6.B reveals that ID-CONNECT outperform other approaches when TPR and candidate set are analyzed together. ID-CONNECT achieve the same TPR as OMF with smaller candidate set. Particularly, ID-CONNECT reduce candidate-set size by more than 50% for any threshold less than 0.6 and achieve similar candidate-set size when the threshold exceeds 0.6.

**2)ER and WS Models:** Figures 6.7 and 6.8 presents results for the ER and WS graph models. The candidate set size of ER model is slightly smaller than that of BA model at small thresholds. This is because of the small clustering coefficient and non-power law degree distribution of the ER model. ID-CONNECT achieves 100% TPR with a small candidate set and OMF achieves similar TPR with comparatively larger candidate-set, as shown in Figure 6.7. The increase in threshold decreases the candidate set size similarly to the BA model.

The TPR and candidate set size for WS model are shown in a Figure 6.8. From Figure 6.8.B it is clear that smaller thresholds yield candidate sets that are larger than candidate sets for ER and BA models; however, at high threshold the size of candidate set for all models is similar. The Watts-Strogatz graph model (WS) exhibits high clustering coefficient and small world network characteristics, along with short average path lengths to most nodes. The small world properties would probably yield comparatively larger candidate set because of high clustering coefficient of nodes.

Analyzing TPR from 6.8.A reveals that TPR of WS is similar to those of the ER and BA models. Comparing approaches for the TPR and candidate set show that ID-

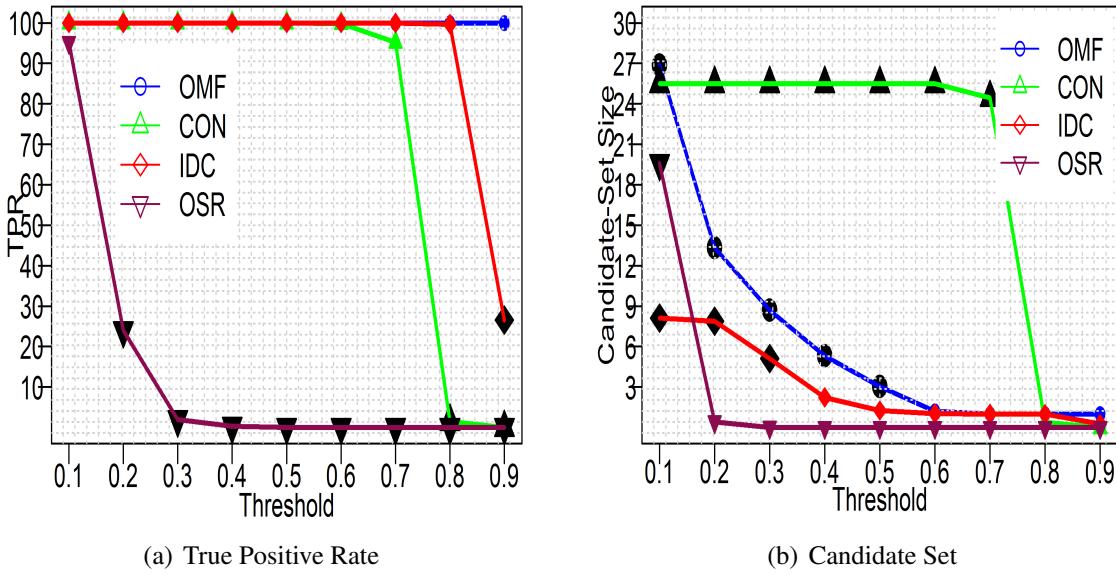


Figure 6.8: Performance Results for Small World Network.

CONNECT and other approaches (especially OMF) achieve 100% TPR at small thresholds but as threshold increases ID-CONNECT has smaller TPR than OMF. From Figures 6.8.A and 6.8.B we can see that OMF achieves 100% TPR at threshold less than .3 with the inclusion of more than 20 identities in the candidate set, and that ID-CONNECT achieves the same TPR with less than 10 identities for the same threshold. The results from Figures 6.8.A and B show that considering edge weights for identity linking in a real network greatly reduces the candidate set size without effecting TPR.

**3) Effect of Friendship Overlap:** In this experiment using the BA model we varied the number of common friends such that two identities belonging to a single physical individual share some friends over two time periods. We assume that the identity at  $T_2$  has an overlap in friendship with the identity at  $T_1$  with different overlap percentages (i.e 80%, 50%, 30%). The call duration and call rate distributions remain the same for both identities in the two time periods. We carried out the analysis for the BA model and for ID-CONNECT and OMF because of their performance in the previous results. The results from figures 6.9 ,6.10 and 6.11 show that the TPR of ID-CONNECT is reasonably high even when number of mutual friends are relatively small. The performance of ID-CONNECT seems to be much better than that of OMF in terms of TPR and candidate set. It is clear that the TPR of ID-CONNECT and OMF decrease with the increase in threshold for any percentage of overlap and OMF is unable to link identities when identities share a small number of mutual friends. For this experiment, we also conclude that ID-CONNECT and OMF would not achieve more than 70% TPR even at small thresholds in conditions when friendship overlap is smaller than 30%.

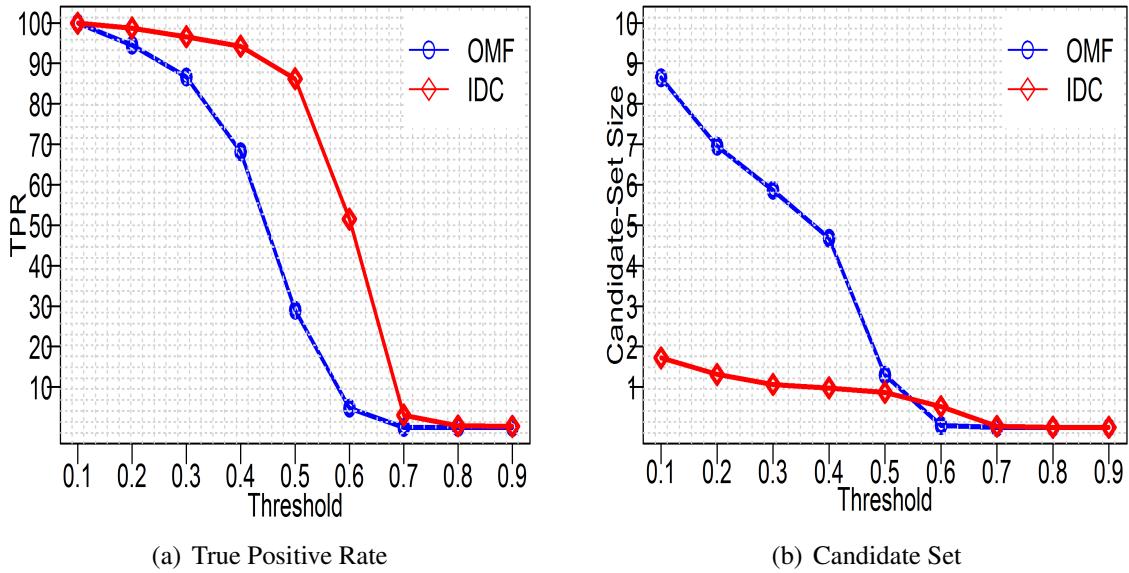


Figure 6.9: Performance Results for 80% Mutual Friends.

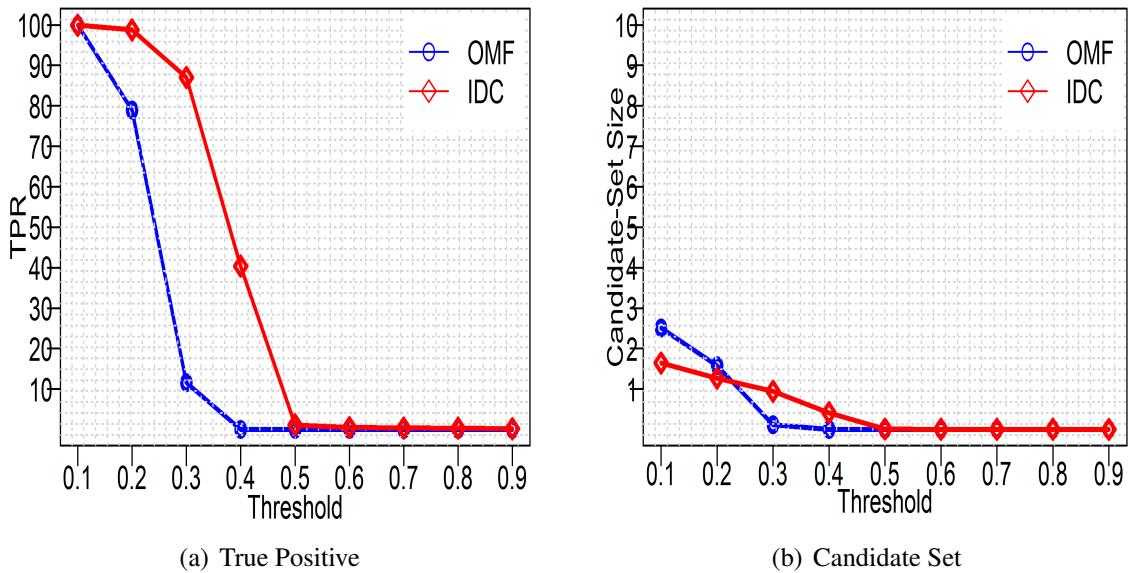


Figure 6.10: Performance Results for 50% Mutual Friends.

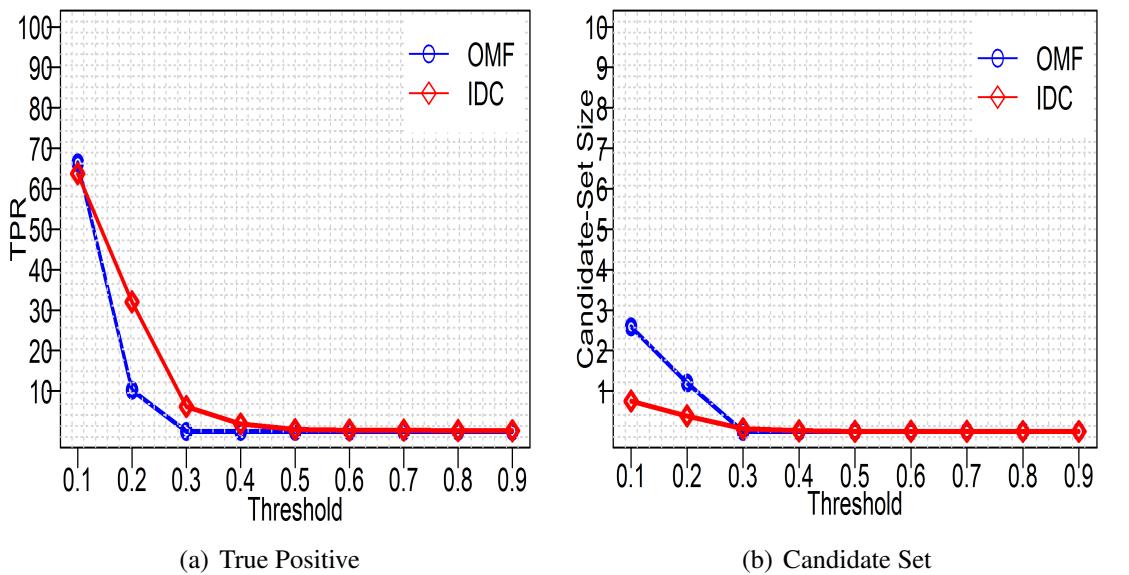


Figure 6.11: Performance Results for 30% Mutual Friends.

**4) Selection of Appropriate Threshold :** The threshold specifies how much overlap in the friendship network and call behavior is required for two compared identities to be considered similar. A high threshold means a high number of mutual friends and a very similar call behavior is needed; a small threshold means that a small number of mutual friends and not a very similar call behavior is required. In the experimental results shown in Figures 6.7, 6.8 and 6.6, we varied the threshold from 0.1 to 0.9. The increase in threshold would not have any effect on TPR until 0.9 but it reduces the candidate-set size. We need to have trade-off between candidate-set size and TPR that is a high TPR and a small candidate-set. From results, we conclude that a threshold value in between 0.3 and 0.5 would provide optimum TPR with a small candidate set and small false positives.

**5) False Positive Rate (FPR):** False Positive Rate (FPR) shows the number of identities that are falsely linked. There is a tradeoff between true positive rate and false positive rate as changing the threshold would likely to affect TPR and FPR simultaneously. The increase in threshold would decreases the FPR and also decreases the TPR. Form the Figure 6.12, it can be seen that IDC has a small FPR (less than 1%) for the BA and ER model. However, the FPR of WS model is very high (greater than 30%) when the threshold is less than 0.4 but as threshold increases the FPR rate is decreased to less than 10% and further decreases to zero as threshold reaches to 1. In WS model, user can share large number of friends because of small world and high clustering coefficient. We can attribute this high false positive to small world and high clustering coefficient of WS model. We particularly adjust the threshold in such a way it should provide small false positive and high true positive. Analyzing FPR and TPR collectively, our results revealed that IDC system is able to have acceptable TPR and FPR for the thresholds between 0.3 and 0.8 for any types of graph network.

### 6.4.2 Application to Spam Detection

In this section, we analyzed the application of identity linking for identifying spammers frequently changing their identities and compare its performance with the system performing spam detection without any identity linking. In this section, we compute reputation of an individual (after linking his identities) and classify individuals as spammer if global reputation score of an individual is less than automated threshold (section 6.3.3).

#### 6.4.2.1 Data-Set for Spam Detection

We generated a synthetic data-set for spam and non-spam users using the approach presented in [AM13] [CMP+13]. However, in this simulation setup we changed identities

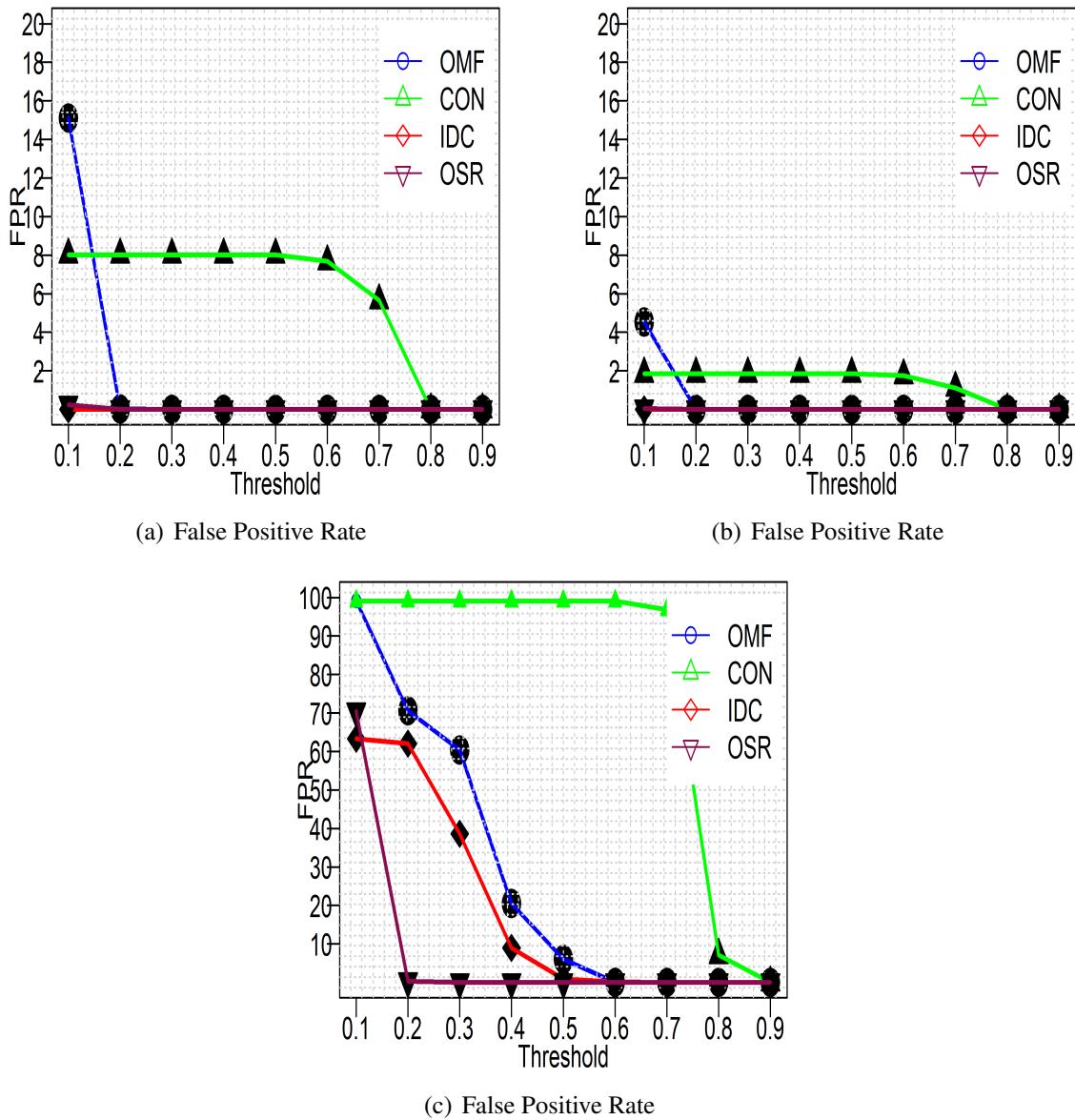


Figure 6.12: False Positive Rate for A) Barabási-Albert, B) Erdős Réenyi, and 3) Small World Networks.

of spammers over the time for different overlaps in a friendship network. We performed simulations for 1000 users, 250 spammers and for the five days.

We use the standard metrics to measure the performances of spam detection system. A legitimate user that is classified as spam by the detection system is termed as a false positive. The false positive rate is defined as the ratio of the number of false positive user to the total number of legitimate users. A spam user that is classified as spammer by the detection system is termed as a true positive. The true positive rate is defined as the ratio of the number of true positive subscriber to the total number of legitimate users. An efficient

detection system is one that achieves high true positive rate and small false positive rate. This can also be characterized by the accuracy of the system which is the ratio of sum of true positive and true negative to the total number of spam and non-spam users. The TPR, FPR and accuracy are computed as  $TPR = TP/(TP+FN)$ ,  $FPR = (FP)/(TN+FP)$  and  $ACC = (TP+TN)/(TP+TN+FN+FP)$ . The evaluation metrics can be explained through the confusion matrix illustrated Table 4.1 of Chapter 4.

#### 6.4.2.2 Spam Filtering Effectiveness

Figure 6.13 presents the true positive rate for different network scenarios over the number of days. We performed analysis for four similar networks of spammers (100%, 80%, 50% and 20% overlap in friendship network) from their different identities. The true positive rate of non-identity linking system decreases over the time this is because the spammer changes his identity as soon as his reputation starts falling. The detection based on identity linking shows improvement in true positive rate as compared to non-identity linking system. Specifically, Figure 6.13 shows that the spammers having greater overlap in friendship network achieves high true positive rate shown in Figure 6.13.A than the spammers having small overlap in friendship network as shown in Figure 6.13.B. This means that in-order to bypass the system the spammer needs not to repeat the identity of victims or repeat only small percentage of identities. The system would only achieve true positive rate of around 60% even in five days. This small true positive is due to the fact that linking identities together would also induce that some spammer would end up with repetitive calling behavior with their target victims which results in high reputation scores. IDC achieves high true positive than the non-linking system for all network scenarios. This is because of the fact that IDC computes the reputation for individual by considering all his identities.

The results from the Figure 6.13 reveals that true positive rate of all system except 100% similar network greatly decreases over the time. This is because of two factors: firstly 25th percentile only detects top spammers and secondly there are many spammers who become reputed because of repetitive calling behavior. The detection rate can be improved blocking the identities identifies as spammers on a particular but use friendship network of block identities for linking it to new identities from next time period.

Figure 6.14 presents the false positive rate for different network scenarios over the number of days. We analyzed FPR for four networks (100%, 80%, 50% and 20% overlap in friendship network). The results from Figure 6.14 reveal that all system achieves zero false positive rates over the time. This means that no legitimate user is mistakenly blocked by the system. We attribute this small false positive rate to the selection of classification

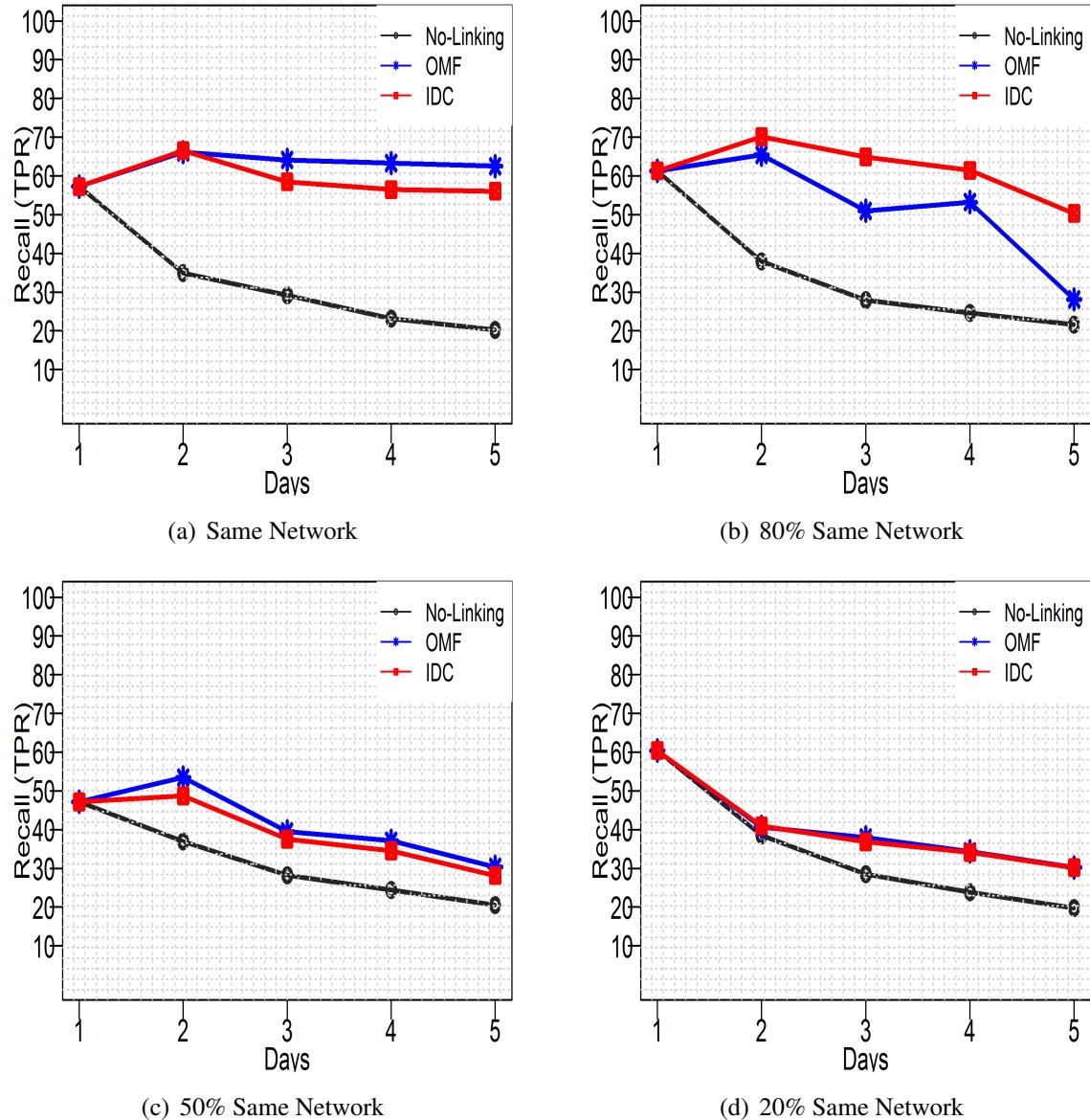


Figure 6.13: True Positive Rate for Different Overlaps in Victim Network.

threshold and a reputation method that collectively use number of social and call features. From Figure 6.14 we can also see that IDC has zero false positive for the network where spammers have small overlap in the target victims, this is because IDC misses linking some of the identities that belongs to the spammers, and individual spamming identities manages reputation score greater than classification threshold.

The accuracy provides a trade-off between correct classification (that classifying spammer as spammer and non-spammer as non-spammer). A high accuracy means the detection system correctly allow non-spammer and block spammer. Figure 6.15 shows accuracy of spam detection system for different network sizes over the number of days. Identity linking based spam detection achieves better detection accuracy than the non-identity linking based system. Specifically identity linking based detection system achieves accuracy of above 85% for highly similar victims of spammer and achieves accuracy of up to 60% over the time for the small percentage of overlapped victims.

## 6.5 Discussion and Limitations of EIS

We presented EIS system that helps in identifying spammers frequently changing their identities by linking similar identities that belongs to one physical individual. The service provider can deploy EIS system as a standalone system or integrate it with call record database or call processing system. In both implementations, the EIS system requires access to CDRs database for the construction of call graphs of identities. The threshold used in EIS system are adjustable and service provider can use its own thresholds for identity linking and spamming classification based on his spam detection policies. The placement of EIS system in a service provider has some privacy concerns but these can be addressed using the approach presented in 4. In identity linking process, EIS can also be used in combination with speech processing engine for validating the identities in a candidate set for a given identity.

The proposed system has been evaluated on the synthetic data-set. Though the evaluation has been performed on all possible network structure but real network graphs are different from the synthetic graphs. The system has following limitations regarding data-set that is handling of a large amount of data, processing of sparse network structure and classification threshold that yields small false positive and high true positive. The other issue that needs to be addressed while performing analyses on real data is to find the ground truth for the identities that belong to one individual. We believe that ID-CONNECT module works well in number of scenarios but it can also link different spammers together if spammers acquire target identities using same technique or acquire identities from the

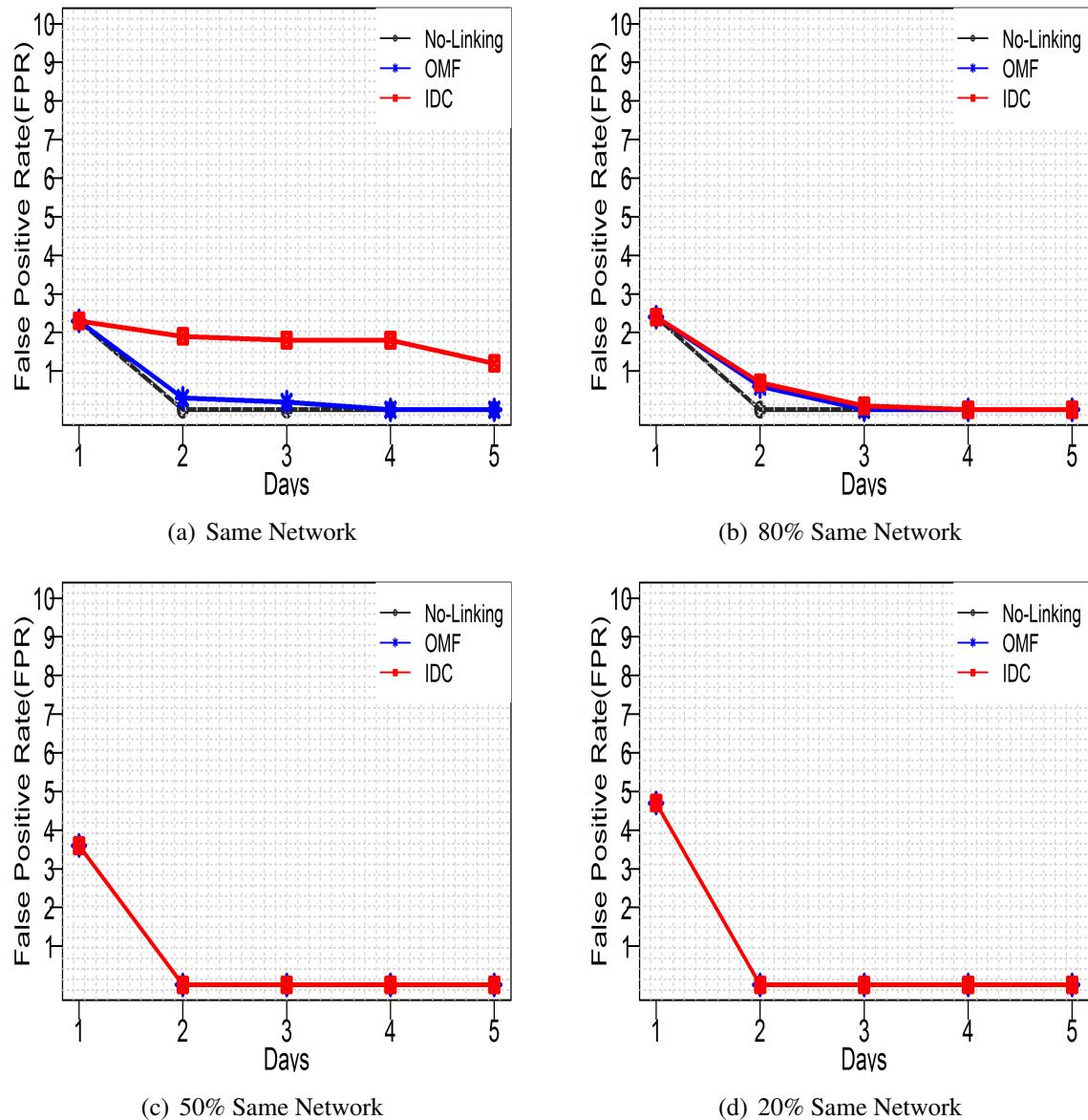


Figure 6.14: False Positive Rate for Different Overlaps in Victim Network.

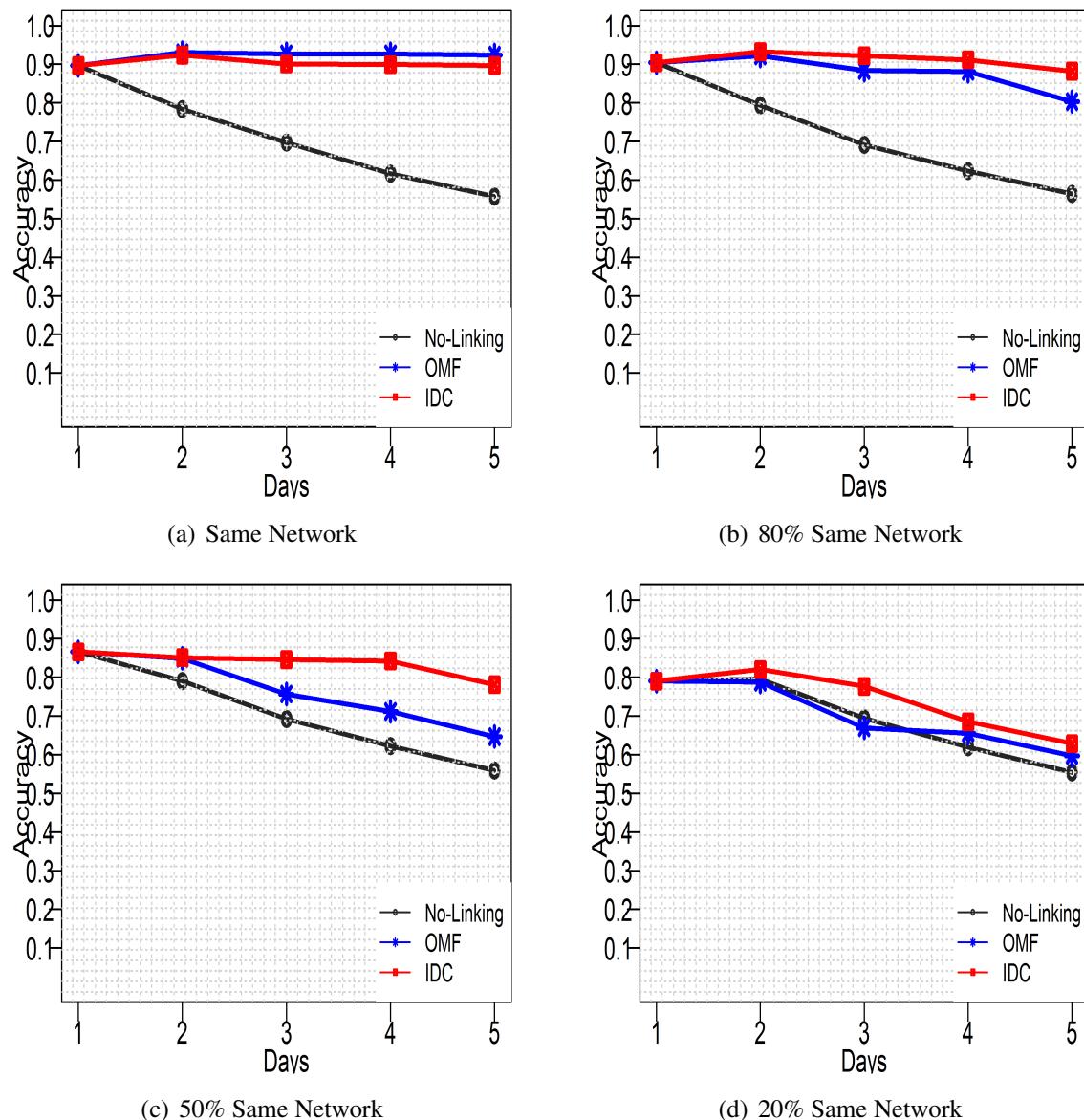


Figure 6.15: Accuracy for Different Overlaps in Victim Network.

same source. Additionally, the EIS system prolonged the detection if spammer target same victims from different identities with different call duration and call rate. Moreover, from results for spam detection, we can see that EIS has small true positive rate when the overlap in victim network of spammer between his identities decreases, but positive side is that EIS has small positive rate in these conditions. We are still investigating ways to address this issue and present a method that not only works for high overlap but also work effectively for the small overlap.

## 6.6 Conclusions

In this chapter, we have presented the design and evaluation of EIS filtering system that decides about behavior of an individual by linking all identities that belong to the individual. Rather than making decision for each identity, EIS system links similar identities that belong to a one physical individual to determine whether an individual is spammer or not. EIS system has three modules: first an ID-CONNECT module—that connects similar identities by estimating the weighted similarity measure between identities using call and social network features. Second, it computes reputation of an individual by using different social network and call features, and thirdly it automatically computes classification threshold below which individuals are considered as spammer. We have evaluated the system on a synthetic data-set for two aspects: identity linking capability and spam detection efficiency. Our evaluation results show that the proposed approach significantly decreases the candidate set size while maintaining a high linking accuracy. In terms of spam detection, the evaluation results show that EIS shows effective resistance against spammers having many identities. Specifically, EIS blocks up to 60% spammers when victim overlap among spammer identity is high and decreases to around 20% when overlap in victim identities is small. We believe that EIS along with content independent speech processing on a candidate set would also significantly improve the identity linking process and achieves high true positive rate.

# Chapter 7

## Conclusions

The research presented in this thesis provides models for blocking unwanted subscribers in a VoIP and voice networks. In this chapter, first we briefly review the contributions made by this thesis, and secondly we briefly provide our thoughts on the future works.

### 7.1 Contributions

VoIP has become a cost effective mechanism for making cheap long distance and international calls. Small and large enterprises are adopting VoIP communication because of its integration with the Internet and cheap calling rates. These benefits of VoIP technology have also attracted telemarketers, prank subscribers, spammers and intruders to exploit the VoIP system for unsolicited spam calls to the subscriber of the technology. These unwanted calls would not only degrade the productivity of humans but could also bring some financial loss. For blocking and limiting these unwanted calls, we presented a spam detection framework that analyzes the behavior of subscriber and incorporate collaboration among service providers for the early and timely identification of spammers. The presented framework involves modeling of behavioral patterns of a subscriber towards his called callees, incorporates collaboration among multiple autonomous service providers for privacy-aware collaboration and allows the linking of similar identities together for early identification of spammers that frequently change their calling identities. Specifically, this thesis makes the following contributions:

**Caller-REP:** We proposed Caller-REP, a system that uses call and social network statistics of the subscriber for the computation of subscriber's global reputation within the service provider network. The global reputation is then compared with the automated threshold for determining if the subscriber is spammer or not. In particular, we provide a system that exploits call duration of subscriber with his called callees in both direction,

call-rate of the subscriber with his called callees in both directions and out-degree distribution of the subscriber for computing direct trust of subscriber with his peer callee. The direct trust scores of the subscriber are then used along with a power iteration algorithm for computing global reputation of the subscriber. Finally, an automated threshold is computed for determining subscriber spamming or non-spamming behavior. We analyzed the performance of Caller-REP system using synthetic data-set that we have generated by simulating the social behavior of spammers and non-spammers. We show that the Caller-REP system achieves a false positive rate of less than 10% and true positive rate of almost 80% in the first two days even in the presence of a significant number of spammers. This increases to a true positive rate of 99% and drops a false positive rate to less than 2% on the third day. In a network with no spammers, our approach achieves a false positive rate of less than 10%. In a network heavily saturated with more than 60% of spam subscribers, our approach achieves a true positive rate of 98% and no false positives. The results show that our approach outperforms other reputation-based detection systems in terms of detection accuracy and detection time.

**COSDS:** Spammers and telemarketers target a very large number of recipients usually dispersed across many Service Providers (SPs). Existing spam detection approaches classify a user as spammer or non-spammer based on the user's Call Detail Records (CDRs) at a single SP. Collaboration and Information sharing between service providers would increase the detection accuracy but detection effectiveness depends on the amount of information shared between service providers. Having service provider's exchange call detail records would arguably attain the best detection accuracy but would require significant network resources. Moreover, service providers are likely to feel uncomfortable in sharing their call records because call records contain user's private information as well as operational details of their network. Having service provider exchange summarized information would reduce network usage and likely be acceptable by SP but could potentially deteriorate the detection accuracy and time. To better understand the effectiveness of collaboration for voice spam detection, we propose COSDS (Collaborative Spit Detection System), a collaborative reputation aggregation and spam detection system for the VoIP network. COSDS uses a trusted Centralized Repository (CR) which computes global reputation (GR) of end-users by aggregating their local reputation scores that are forwarded by the collaborating SPs. The local reputation scores transferred to the CR by the collaborating SPs are less demanding in terms of network resources and do not contain call records. We evaluate our system using synthetic data that we generate through a model of spammer and non-spammer social behavior and in a network with five collaborating SPs. The results show that the COSDS approach has better detection accuracy than the traditional stand-alone detection systems. For spammers making calls to recipients of

many SPs, COSDS has True Positive (TP) rate of 80% and False Positive (FP) rate of 2% on a first day which further increases to 100% TP rate with zero FP rate in three days. Subject to moderate spamming rate, COSDS is able to suppress all spammers in five days with a FP rate of less than .5%. The results also show that COSDS detection accuracy is comparable to that of a system where collaboration is achieved through the exchange of call records. COSDS approach is fast, requires much less communication overhead and only requires a few iterations for reputation convergence within the SP.

**EIS:** Multiple identities are created to gain financial benefits by performing malicious activities such as spamming, committing frauds and abusing the system. A single malicious individual may have a large number of identities in order to make malicious activities to a large number of legitimate individuals. Linking identities of an individual would help in protecting the legitimate users from abuses, frauds, and maintains reputation of the service provider. Simply analyzing each identity's historical behavior is not sufficient to block spammer frequently changing identity because spammer quickly discards the identity and start using new one. Moreover, spammers may appear as a legitimate user on an initial analysis, for example because of small number of interactions from any identity. The challenge is to identify the spammer by analyzing the aggregate behavior of an individual rather than that of a single calling identity. This contribution presents EIS (Early Identification of Spammers) system for the early identification of spammers frequently changing identities. Specifically, EIS system consists of three modules and uses social call graph among identities. 1) An ID-CONNECT module that links identities that belongs to a one physical individual based on a social network structure and calling attributes of identities; 2) a reputation module that computes reputation of an individual by considering his aggregate behavior from his different identities; and 3) a detection module that computes automated threshold below which individuals are classified as a spammer or a non-spammer. We evaluate the proposed system on a synthetic data-set that has been generated for the different graph networks and different percentage of spammers. Performance analysis shows that EIS is effective against spammers frequently changing their identities and is able to achieve high true positive rate when spammers have high small overlap in target victims from their identities.

Additionally, this thesis also made following minor contributions:

**Privacy Analysis of CDR's:** The call detail records normally have enough information that can be used to infer the relationship network of target identities. The service providers normally share anonymized data to the research organizations for extracting meaningful information for business purposes. However, they only anonymized identities of the subscriber and the called callees leaving call duration and call time as it is in a non-anonymized form. In chapter 5, we have shown that call duration and call time

information can be collectively used to identify the anonymized identity of the users that further can be exploited to infer the relationship network of these users. To overcome this limitation, we proposed that call-time and call duration should also be anonymized in such a way that it would not affect the accuracy of the detection system and intruders would not be able to learn the relationship network of the target user. We experimentally proved that anonymizing call-time and call duration would further minimize the privacy breach through collective use of call features and would return large candidate set against the intruder query for the target victim.

**Data-Set:** To-date no call detailed data set is publicly available for evaluating the performance of detection system. Additionally, service providers are not willing to share information about the spam attacks or call records because of privacy concerns and sensitive information. Due to these issues, we synthetically generated the data-set using social calling behavior of the spammer and non-spammer. To this extent, we have used different graph models for generating the call graph network, used different random distributions for generating the call-rate and call duration of user with friends and non-friends.

## 7.2 Future Works

Though results presented in this thesis have demonstrated the effectiveness against spammers in a VoIP and voice network; however, there are several directions in which the proposed models can be extended.

**Identification of Other Call Features:** In this thesis, we have used call and social network features for computing global reputation of the subscribers within the telecommunication network. These features have shown great effectiveness in timely detection of spammers. However, the presented work can be extended to utilize the call statistics information such as call release cause codes, inter-arrival time between call requests made by the subscriber, number of calls disconnected by the subscriber making calls and number of calls disconnected by the callee subscriber, number of failed calls, etc. We want to integrate these call features along with the Caller-REP system in order to understand the impact of these features on the detection performance and to minimize the false positive rate. We would also like to investigate the ways for personalized spam detection where the spam calls are only allowed to the specific subscribers those wish to receive these spam and promotion calls.

**Secure Multi-party Computation:** In this thesis, we have relied on the use of centralized repository for reputation aggregation and decision about behavior of the subscriber from collaborating service providers. However, the centralized repository can be a single

point of failure, is not scalable, and service providers might not willing to collaborate with the centralized repository. Instead of this collaboration, the service provider might exchange encrypted scores with the peer collaborating service providers. We would like to explore the idea of distributed collaboration where each service provider directly collaborates with his peer service provider without worrying about privacy of his customers. This collaboration approach has two challenges: 1) privacy protection of collaborators data, and 2) convergence and communication overhead required for reputation aggregation. To achieve these objectives, we are interested in a design of Secure Multi-party Computation (SMC) scheme without overwhelming the network bandwidth required for the collaboration process.

**Collaboration Among Different Mobile Applications** The popularity of smart phones has attracted large number of people to replace their features phones with the smart phones and use newly developed VoIP applications for the free voice conversation with friends and family. There are several free VoIP calling applications are in the market for example WhatsApp, Viber and Skype are some of the famous one. The free calling and easy integration of these applications with the spamming tools have also attracted spammers and scammers to target subscribers of these applications with the unwanted calls and messages. Mostly, these applications operate using mobile number of subscriber except for Skype that uses email identity as well. A spammer can exploit subscribers of these applications in parallel or one by one. However, the effect of spamming could be minimized with the intra-application collaboration. We would like to extend this work towards design of such spam detection application that collectively uses information from several VoIP applications and aggregate characterize the behavior of the subscriber targeting an application subscriber. We have already developed an android mobile application as part of master thesis [CAR15] for detecting spammers on a standalone android mobile device using subscriber's call logs and collaboration among end-users. The design of intra-application collaboration has a challenge of developing a common application programming interface for excessing the data from different applications.

**Caller ID Spoofing** Caller ID spoofing is when someone change the Caller ID to any number that he wants to display to the call recipients. Caller ID is actually meant for providing information to callee about the person calling the callee. There are several techniques and smart phone applications are available that can be used to spoof the identity. It can be used to identify the telemarketers and spammers; however, spammers, telemarketers and prank callers are spoofing identities of legitimate users to circumvent the detection system and fraud user by pretending to be a legitimate entity. It is important to have a system that can identify the users hiding their true identity. The challenge in this regard is to two-folds. 1) Making decision during the call setup phase, and 2) making

decision about identity of the caller without involving caller and the callee for a certain response messages. The call setup messages can provide complete information about caller location, calling identity, devices etc. and can be used to block identity spoofers.

# Bibliography

- [ACF13] C.G. Akcora, B. Carminati, and E. Ferrari. "User Similarities on Social Networks", *Social Network Analysis and Mining*, Pages:475–495, 2013.
- [AM11] M.A. Azad and R. Morla. "Multistage SPIT Detection in Transit VoIP", In *19th IEEE SoftCOM*, 2011.
- [AM12] M.A. Azad and R. Morla. "Mitigating SPIT with Social Strength", In *Proceedings of 2012 IEEE TrustCom*, 2012.
- [AM13] M.A. Azad and R. Morla. "Caller-Rep: Detecting Unwanted Calls with Caller Social Strength", In *Computers & Security*, Volume 39 Part B Pages:219–236, 2013.
- [AMF10] L. Akoglu, M. McGlohon, and C. Faloutsos. oddball: Spotting Anomalies in Weighted", In *Preceedings of 14th PAKDD, Book Chapter in Advances in Knowledge Discovery and Data Mining*, 2010.
- [BAP07] V.A. Balasubramaniyan, M. Ahamad, and H. Park. "CallRank: Combating SPIT Using Call Duration, Social Networks and Global Reputation", In *Proceedings of 4th CEAS*, 2007.
- [BGG+16] M. Balduzzi, P. Gupta, L. Gu, Gao.D, and M. Ahamad. "MobiPot: Understanding Mobile Telephony Threats with Honeycards", In *Proceedings of 11th ACM ASIACCS*, 2016.
- [BGH+04] V.D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P.Van Dooren. "A Measure of Similarity Between Graph Vertices: Applications to Synonym Extraction and Web Searching", *SIAM Review*, Volume 46 Pages:647–666, 2004.
- [BKE+13] M. Berlingero, D. Koutra, T. Eliassi-Rad, and C. Faloutsos. "Network Similarity via Multiple Social Theories", In *Proceedings of 2013 ASONAM*, 2013.

- [BKP12] S. Bartunov, A. Korshunov, and S. Park. "Joint Link-Attribute User Identity Resolution in Online Social Networks", In *Proceedings of 6th SNA-KDD Workshop*, August,2012.
- [BSG+11] H.Kaffash. Bokharaei, A. Sahraei, Y. Ganjali, R. Keralapura, and A. Nucci. "You can SPIT, but you can't hide: Spammer Identification in Telephony Networks", In *Proceedings of IEEE INFOCOM*, 2011.
- [BR05] P. Oscar Boykin and Vwani P. Roychowdhury. "Leveraging social networks to fight spam", *Computer*, Volume 38 Pages:61–68, 2005.
- [CAR15] R. Cardoso. "Mobile Application for Blocking Spam Callers", Master's thesis, Faculty of Engineering, University of Porto, 2015.
- [CDN05] P-Alexandru. Chirita, J. Diederich, and W. Nejdl. "MailRank: Using Ranking for Spam Detection", In *Proceedings of 14th ACM CIKM*, 2005.
- [CHA2015] B. Chandrasekaran. "Survey of network traffic models", Retrieved from [http://www.cse.wustl.edu/~jain/cse567-06/ftp/traffic\\_models3.pdf](http://www.cse.wustl.edu/~jain/cse567-06/ftp/traffic_models3.pdf)
- [CJE15] Criminal and Justice Unit England. "Sending Unsolicited ext Messages",
- [CIS18] CISCO Visual Networking Index Service Adoption Forecast 2013 2018 White Paper", Technical report, Cisco, 2013-2018. Retrieved from <http://goo.gl/X3efVF>",
- [CKK05] S. Clauß, D. Kesdogan, and T. Kölsch. "Privacy Enhancing Identity Management: Protection against re-identification and Profiling", In *Proceedings of 2005 workshop on Digital identity management*, 2005.
- [CKV+02] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M.. Zhu. Tools for privacy preserving distributed data mining. *SIGKDD Explor. Newslett.*, Volume 4 Pages:28–34, 2002.
- [CLO16] CloudMark. "CloudMark SpamNet, 2016. Retrieved from <https://www.cloudmark.com/en>",
- [CMP+13] S. Chiappetta, C. Mazzariello, R. Presta, and Simon P. Romano. "An Anomaly-based Approach to the Analysis of the Social Behavior of VoIP Users", *Computer Networks*, Volume 57:1545–1559, 2013.

- [CO05] N.J. Croft and M.S. Olivie. A Model for Spam Prevention in IP Telephony Networks Using Anonymous Verifying Authorities. In *Proceedings of ISSA 2005*, 2005.
- [COB+11] N. Chaisamran, T. Okuda, G. Blanc, and S. Yamaguchi. "Trust-Based VoIP Spam Detection Based on Call Duration and Human Relationships", *IEEE/IPSJ International Symposium on Applications and the Internet*, 2011.
- [DDT+09] S. Dritsas, V. Dritsou, V. Tsoumas, P. Constantopoulos, and D. Gritzalis. On-tospit: {SPIT} management through ontologies. *Computer Communications*, Volume 32(1) Pages:203 – 212, 2009.
- [DK05] R. Dantu and P. Kolan. "Detecting Spam in VoIP Networks. In *Proceedings of The Steps to Reducing Unwanted Traffic on the Internet*, 2005.
- [DSN09] N. D'Heureuse, J. Seedorf, and S. Niccolini. "A Policy Framework for Personalized and Role-based SPIT Prevention", In *Proceedings of 3rd IPTComm*, 2009.
- [DTN11] N. D'Heureuse, S. Tartarelli, and S. Niccolini. "Analyzing Telemarketer Behavior in Massive Telecom Data Records. In *Proceedings of Trustworthy Internet*", Springer, 2011.
- [DVP+04] E. Damiani, S. Vimercati, S. Paraboschi, and P. Samarati. "P2P-based Collaborative Spam Detection and Filtering", In *Proceedings of 4th International Conference on Peer-to-Peer Computing*, 2004.
- [EGM10] S. Ehlert, D. Geneiatakis, and T. Magedanz. "Survey Of Network Security Systems To Counter SIP-Based Denial-Of-Service Attacks", *Computers & Security*, Volume 29(2) Pages:225–243, 2010.
- [FAC14] There are more than 90 million duplicate accounts on facebook", 2014. Retrieved from <http://goo.gl/ozLQ3J>
- [FFF99] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, 1999.
- [HS08] C.A. Hidalgo and C. Rodriguez Sickert. "The dynamics of a mobile phone network", *Physica A: Statistical Mechanics and its Applications*, Volume 387 Pages:3017–3024, 2008.

- [FZN06] N. Foukia, L. Zhou, and C. Neuman. "Multilateral Decisions for Collaborative Defense Against Unsolicited Bulk e-MailMul", In *Proceedings of 4th International Conference on Trust Management*, 2006.
- [FZX15] H. Fu, A. Zhang, and X. Xie. "Effective Social Graph Deanonymization Based on Graph Structure and Descriptive Information", *ACM Transactions on Intelligent Systems and Technology*, Volume 6 pages:1–49:29, 2015.
- [GCA+04] Luiz Henrique Gomes, Cristiano Cazita, Jussara M. Almeida, Virgílio Almeida, and Wagner Meira, Jr. "Characterizing a spam traffic", In *Proceedings of the 4th ACM IMC*, 2004.
- [GKB+12] D. Gritzalis, P. Katsaros, S. Basagiannis, and Y. Soutoupis. "Formal Analysis for Robust anti-SPIT Protection using Model Checking", *International Journal of Information Security*, Volume 11 Pages:121–135, 2012.
- [GKT+10] Del. Genio, H. Kim, Z. Toroczkai, and E. Bassler. "Efficient and Exact Sampling of Simple Graphs with Given Arbitrary Degree Sequence", *PLoS ONE*, Volume 5:e10012, 2010.
- [GLP+13] O. Goga, H. Lei, Sree Hari K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. "Exploiting Innocuous Activity for Correlating Users Across Sites", In *Proceedings of 22nd WWW*, 2013.
- [GM08] D. Gritzalis and Y. Mallios. A sip-oriented {SPIT} management framework. *Computers & Security*, Volume 27 Pages:136–153, 2008.
- [GRA73] M.S. Granovetter. "The Strength of Weak Ties", In *Proceedings of The American Journal of Sociology*, pages:1360–1380. JSTOR, 1973.
- [GSB+15] P. Gupta, B. Srinivasan, V. Balasubramaniyan, and M. Ahamed. "Phoneypot: Data-driven Understanding of Telephony Threats", In *Proceedings of 20th NDSS*, 2015.
- [GTM+14] N. Z. Gong, A. Talwalkar, L. Mackey, L. Huang, E. Chul. Shin, E. Stefanov, E. Shi, and D. Song. "Joint Link Prediction and Attribute Inference Using a Social-Attribute Network", *ACM Transactions on Intelligent Systems Technology*, Pages:1–27, 2014.
- [GYH08] H. Guang, W. Ying, and Z. Hong. "SPIT Detection and Prevention Method Based on Signal Analysis", In *Proceedings of 3rd International Conference on Convergence and Hybrid Information Technology*, 2008.

- [HGL+11] K. Henderson, B. Gallagher, L. Li, L. Akoglu, T. Eliassi-Rad, H. Tong, and C. Faloutsos. "It's Who You Know: Graph Mining Using Recursive Structural Features", In *Proceedings of 17th ACM SIGKDD*, 2011.
- [HHM+06] M. Hansen, M. Hansen, J. Möller, T. Rohwer, C. Tolkmitt, and H. Waack. "Developing a Legally Compliant Reachability Management System as a Countermeasure against SPIT", In *Proceedings of 3rd Annual VoIP Security Workshop*, 2006.
- [HSZ+06] H. Hong, K. Sripanidkulchai, H. Zhang, Z. Shae, and D. Saha. "Incorporating Active Fingerprinting into SPIT Prevention Systems", In *Proceedings of 2006 VSW*, 2006.
- [IFA+11] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. "Identifying Users Across Social Tagging Systems", In *Proceedings of 2011 ICWSM*, 2011.
- [INF15] Infonetics Research for VoIP Services Forecast (last retrieved April 2016), 2015. Retrieved from <http://goo.gl/rh21kQ>,
- [ISW13] SA. Iranmanesh, H. Sengar, and H. Wang. "A Voice Spam Filter to Clean Subscribers Mailbox", In *Proceedings of Springer Security and Privacy in Communication Networks*, Pages: 349–367, 2013.
- [ITU15] ITU Survey On Anti-spam Legislation Worldwide, July 2015.
- [JEN15] C. KERR JENNIFER. "Complaints about Automated Calls up Sharply (last retrieved August 2015). Retrieved from <http://goo.gl/H5HTBh>",
- [JKJ13] P. Jain, P. Kumaraguru, and A. Joshi. "@I Seek 'Fb.Me': Identifying Users Across Multiple Online Social Networks", In *Proceedings of 22nd WWW*, 2013.
- [JMG+09] John P. John, A. Moshchuk, Steven D. Gribble, and A. Krishnamurthy. "Studying spamming botnets using botlab", In *Proceedings of the 6th NSDI'09*, 2009.
- [JMN+13] T. Jung, S. Martin, M. Nassar, D. Ernst, and G. Leduc. "Outbound SPIT Filter with Optimal Performance Guarantees", *Computer Networks*, Volume 57 Pages:1630-1643, 2013.
- [JTJ11] Lei Jin, Hassan Takabi, and James B.D. Joshi. "Towards active detection of identity clone attacks on online social networks", In *Proceedings of the First CODASPY*, 2011.

- [JW02] G. Jeh and J. Widom. "SimRank: A Measure of Structural-context Similarity", In *Proceedings of 8th ACM SIGKDD*, 2002.
- [KD07] P. Kolan and R. Dantu. "Socio-Technical Defense Against Voice Spamming", *ACM Transactions on Autonomous and Adaptive Systems*, Volume 2, 2007.
- [KER11] A. Keromytis. "A Comprehensive Survey of Voice over IP Security Research", *IEEE Communications Surveys Tutorials*, Pages:1–24, 2011.
- [KJF+06] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi. "Detecting Spam Blogs: A Machine Learning Approach", In *Proceedings of 21st National Conference on Artificial Intelligence*, 2006.
- [KRS+06] J.S. Kong, B.A. Rezaei, N. Sarshar, V.P. Roychowdhury, and P.O. Boykin. "Collaborative Spam Filtering Using E-Mail Networks", *Computer*, Volume 39 Pages:67–73, 2006.
- [KSM03] D. Kamvar, T. Schlosser, and H. Molina. "The Eigentrust Algorithm for Reputation Management in P2P Networks", In *Proceedings of 12th WWW*, 2003.
- [LAW15] Canada's anti-spam Legislation (CASL), 2015. Retrieved from <http://goo.gl/8vwZTJ>,
- [LCL+15] J. Lee, K. Cho, C. Lee, and S. Kim. "VoIP-aware Network Attack Detection based on Statistics and Behavior of SIP Traffic", *Peer-to-Peer Networking and Applications*, 2015.
- [LCW10] K. Lee, J. Caverlee, and S. Webb. "Uncovering Social Spammers: Social Honeypots + Machine Learning", In *Proceedings of 33rd ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010.
- [LGK+11] D. Lentzen, G. Grutzek, H. Knospe, and C. Porschmann. "Content-Based Detection and Prevention of Spam over IP Telephony - System Design, Prototype and First Results", In *Proceedings of 2011 IEEE ICC*, 2011.
- [LK07] J. Lindqvist and M. Komu. "Cure for Spam Over Internet Telephony", In *Proceedings of 4th IEEE CCNC*, 2007.
- [LWZ+14] S. Liu, S. Wang, F. Zhu, and Zhang. "HYDRA: Large-scale Social Identity Linkage via Heterogeneous Behavior Modeling", In *Proceedings of 2014 ACM SIGMOD*, 2014.

- [LY07] H. Lam and D. Yeung. A Learning Approach to Spam Detection Based on Social Networks. In *Proceedings of 4th CEAS*, 2007.
- [LZR09] K. Li, Z. Zhong, and L. Ramaswamy. "Privacy-Aware Collaborative Spam Filtering", *IEEE Transactions on Parallel and Distributed Systems*, Volume 20 Pages:725–739, 2009.
- [MLG+06] B. Mathieu, Q. Loudier, Y. Gourhant, F. Bougant, and M. Osty. "SPIT Mitigation by a Network-Level Anti-Spit Entity", In *Proceedings of 3rd Workshop on Securing Voice over IP*, June 2006.
- [MLM+12] A. Malhotra, C.T. Luam, W. Jr. Meira, P. Kumaraguru, and V. Almeida. "Studying User Footprints in Different Online Social Networks", In *Proceedings of 12th ASONAM*, 2012.
- [LY07] H. Lam and D. Yeung. "A learning approach to spam detection based on social networks", In *Proceedings of CEAS*, 2007.
- [MMG+07] A. Mislove, M. Marcon, P. Gummadi, Krishna, P. Druschel, and B. Bhattacharjee. "Measurement and Analysis of Online Social Networks", In *Proceedings of 7th ACM SIGCOMM Conference on Internet Measurement*, 2007.
- [MMR02] S. Melnik, H. Molina, and E. Rahm. "Similarity Flooding: A Versatile Graph Matching Algorithm And Its Application To Schema Matching", In *Proceedings of 18th Data Engineering*, 2002.
- [MNS08] B. Mathieu, S. Niccolini, and D. Sisalem. "SDRS: A Voice-over-IP Spam Detection and Reaction System", *IEEE Security Privacy*, 6:52–59, 2008.
- [MOT12] F. Moradi, T. Olovsson, and P. Tsigas. "Towards Modeling Legitimate and Unsolicited Email Traffic Using Social Network Properties", In *Proceedings of 5th Workshop on Social Network Systems*, 2012.
- [MV05] R. MacIntosh and D. Vinokurov. "Detection and Mitigation of Spam in IP Telephony Networks using Signaling Protocol Analysis", In *Proceedings of 2005 IEEE/Sarnoff Symposium on Advances in Wired and Wireless Communication*, 2005.
- [NCM12] A. Nunes, P. Calado, and B. Martins. "Resolving User Identities over Social Networks Through Supervised Learning and Rich Similarity Features", In *Proceedings of 27th Annual ACM Symposium on Applied Computing*, 2012.

- [NEW05a] M. E. J. Newman. "Power laws, Pareto Distributions and Zipfs Law", In *Proceedings of Contemporary Physics*, Pages:323–351. 2005.
- [NEW05b] M. E. J. Newman. "Random graphs as Models of Networks", Pages:35–68, 2005.
- [NEW10] M. E. J. Newman. *Networks An Introduction*. Oxford University Press, 2010.
- [NGD+06] A.A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. "On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications", In *Proceedings of 15th ACM CIKM '06*, 2006.
- [NNM+06] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. "Detecting Spam Web Pages Through Content Analysis", In *Proceedings of 15th International WWW*, 2006.
- [NNS+07] M. Nassar, S. Niccolini, R. State, and T. Ewald. "Holistic VoIP Intrusion Detection and Prevention System", In *Proceedings of 1st IPTCOMM*, 2007.
- [NS09] A. Narayanan and V. Shmatikov. "De-anonymizing Social Networks", In *Proceedings of 30th IEEE Symposium on Security and Privacy*, 2009.
- [NSC+08] A.A. Nanavati, R. Singh, D. Chakraborty, K. Dasgupta, S. Mukherjea, G. Das, S. Gurumurthy, and A. Joshi. "Analyzing the Structure and Evolution of Massive Telecom Graphs", *IEEE Transactions on Knowledge and Data Engineering*, Volume 20 Pages:703–718, 2008.
- [PIN16] "The State of Phone Fraud 2014-2015 a Global, Cross-Industry",2016. Retrieved from <https://www.pindrop.com/phone-fraud-report/>.
- [OS09] K. Ono and H. Schulzrinne. "Have I met you before?: using Cross-Media Relations to Reduce SPIT", In *Proceedings of 3rd IPTCOMM*, 2009.
- [OZ04] S. Oliveira and O.R. Zaïane. "Toward Standardization in Privacy-Preserving Data Mining", In *Proceedings of Data Mining Standards (DM-SSP 2004), conjuction with KDD 2004*, 2004.
- [PAI99] P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *Proceedings of EUROCRYPT 99 ,Springer Lecture Notes in Computer Science* , 1999.

- [PD09] S. Phithakkitnukoon and R. Dantu. "Defense against SPIT using Community Signals", In *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, 2009.
- [PEN04] J. Penders. "Privacy in (mobile) Telecommunications Services", Kluwer Academic Publishers, 2004 pages:247–260. .
- [PER09] SARAH PEREZ. "Who Uses Social Networks and What Are They Like ((last Retrieved August 2015), JUL 2009.
- [PGK+08] P. Patankar, N Gunwoo, G. Kesidis, and C.R. Das. "Exploring Anti-Spam Models in Large Scale VoIP Systems", In *Proceedings of 28th International Conference on Distributed Computing Systems*, 2008.
- [PK08] C. Pörschmann and H. Knospe. "Analysis of Spectral Parameters of Audio Signals for the Identification of Spam Over IP Telephony", In *Proceedings of 5th CEAS*, 2008.
- [QNT+07] J. Quittek, S. Niccolini, S. Tartarelli, M. Stiemerling, M. Brunner, and T. Ewald. "Detecting SPIT Calls by Checking Human Communication Patterns", In *Proceedings of IEEE ICC*, 2007.
- [QNT+08] J. Quittek, S. Niccolini, S. Tartarelli, and R. Schlegel. "On Spam over Internet Telephony (SPIT) Prevention", In *Proceedings of IEEE Communications Magazine*, Pages:80–86. 2008.
- [RCD10] E. Raad, R. Chbeir, and A. Dipanda. "User Profile Matching in Social Networks", In *Proceedings of 13th NBiS*, pages 297–304, 2010.
- [REP15] Symantec Intelligence Report (Retrieved September 2015), 2015. Retrieved from <https://goo.gl/4C7xu5>",
- [RFC3261] J. Rosenberg, H. Schulzrinne, G. Camarillo, P. J. Johnston, R. Sparks, M. Handley, and E. Schooler, " Sip: Session initiation protocol. IETF RFC 3261",2002.
- [RFV07] A. Ramachandran, N. Feamster, and S. Vempala. "Filtering Spam with Behavioral Blacklisting", In *Proceedings of 14th ACM CCS*, 2007.
- [RJ08] J Rosenberg and C Jennings. The session initiation protocol (sip) and spam (rfc 5039), 2008.

- [RS05] Y. Rebahi and D. Sisalem. "SIP Service Providers and the Spam Problem", In *Proceedings of Voice over IP Security Workshop*, 2005.
- [RSM06] Y Rebahi, D Sisalem, and T. Magedanz. "SIP Spam Detection", In *Proceedings of ICDT '06*, 2006.
- [SAS06] D. Shin, J. Ahn, and C Shim. "Progressive Multi Gray-Leveling: a Voice Spam Protection Algorithm", In *Proceedings of IEEE Network*, Pages:18–24. 2006.
- [SBK07] G. Singaraju and B. ByungHoon-Kang. "RepuScore: Collaborative Reputation Management Framework for Email Infrastructure", In *Proceedings of 21st Conference on Large Installation System Administration Conference*, 2007.
- [SCA08] M. Szomszor, I. Cantador, and H. Alani. "Correlating User Profiles from Multiple Folksonomies", In *Proceedings of 9th ACM Conference on Hypertext and Hypermedia*, 2008.
- [SDG08] Y. Soupionis, S. Dritsas, and D. Gritzalis. "An Adaptive Policy-Based Approach to SPIT Management", In *Proceedings of ESORICS Computer Security*, 2008.
- [SDN+09] J. Seedorf, N. D'Heureuse, S. Niccolini, and M. Cornolti. "Detecting Trustworthy Real-Time Communications Using a Web-of-Trust", In *Proceedings of IEEE GLOBECOM*, 2009.
- [SG10] Y. Soupionis and D. Gritzalis. "Audio CAPTCHA: Existing Solutions Assessment and a New Implementation for VoIP Telephony", *Computer & Security*, Volume 29 Pages:603–618, 2010.
- [SKE+14] Y. Soupionis, R. Koutsiamanis, P. Efraimidis, and D. Gritzalis. "A Game-theoretic Analysis of Preventing Spam over Internet Telephony via Audio CAPTCHA-based Authentication", *Journal Computer Security*, Volume 22 Pages:383–413, 2014.
- [SKY11] M. Sirivianos, K. Kyungbaek, and X. Yang. "SocialFilter: Introducing Social Trust to Collaborative Spam Mitigation", In *Proceedings of IEEE INFOCOM*, 2011.
- [SLS+10] A.U. Schmidt, A. Leicher, Y. Shah, Inhyok Cha, and L. Guccione. "Sender Scorecards for the Prevention of Unsolicited Communication", In *Proceedings of 2nd IEEE Workshop on Collaborative Security Technologies*, 2010.

- [SMG+12] J. Strobl, B. Mainka, G. Grutzek, and H. Knospe. "An Efficient Search Method for the Content-based Identification of Telephone-SPAM", In *Proceedings of 2012 IEEE ICC*, 2012.
- [SML+14] I. Santos, I. Marcos, C. Laorden, P. García, A. Ibirika, and P. Bringas. "Twitter Content-Based Spam Filtering", In *Proceedings of International Joint Conference SOCO13-CISIS13-ICEUTE13*, 2014.
- [SMS+08] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove. "Mobile call graphs: beyond power-law and lognormal distributions", In *Proceedings of 14th ACM SIGKDD*, 2008.
- [SNT+06] R. Schlegel, S. Niccolini, S. Tartarelli, and M. Brunner. "SPam over Internet Telephony (SPIT) Prevention Framework", In *Proceedings of IEEE GLOBECOM*, 2006.
- [SPA2015] Spam Phone Calls Cost U.S. Small Businesses Half-Billion Dollars in Lost Productivity, Marchex Study Finds", Retrieved from <http://goo.gl/jTrgp3>",
- [SS04] K. Srivastava and H. Schulzrinne. "Preventing Spam for SIP-based Instant Messages and Sessions", Technical report, Columbia University Technical Report CUCS-042-04, October 2004.
- [SS09] C. Sorge and J. Seedorf. "A Provider-Level Reputation System for Assessing the Quality of SPIT Mitigation Algorithms", In *Proceedings of IEEE ICC*, 2009.
- [TWI16] "36% of tweets contain links", Retrieved from <http://goo.gl/UG6c4R>.
- [SSB05] E. Spertus, M. Sahami, and O. Buyukkokten. "Evaluating Similarity Measures: A Large-scale Study in the Orkut Social Network", In *Proceedings of 11th ACM SIGKDD*, 2005.
- [3GPP2015] Study of Mechanisms for Protection against Unsolicited Communication for IMS (PUCI).
- [IMS2015] Study of Mechanisms for Protection against Unsolicited Communication for IMS (PUCI). 3GPP Standard 2015.
- [STM] H Schulzrinne, H Tschofenig, and J Morris. "Common Policy: A Document Format for Expressing Privacy Preferences (RFC 4745).

- [SV99] J.A.K. Suykens and J. Vandewalle. "Least Squares Support Vector Machine Classifiers", *Neural Processing Letters*, Volume 9 Pages:293–300, 1999.
- [SWN12] H. Sengar, X. Wang, and A. Nichols. "Call Behavioral Analysis to Thwart SPIT Attacks on VoIP Networks", In *Proceedings of Security and Privacy in Communication Networks*, Pages:501–510, 2012.
- [SXB11] H. Sengar, W. Xinyuan, and A. Nichols. "Thwarting Spam over Internet Telephony (SPIT) Attacks on VoIP Networks", In *Proceedings of 19th IEEE Quality of Service*, 2011.
- [TDZ+16] H. Tu, A. Doupé, Z. Zhao, and G. Ahn. "SoK: Everyone Hates Robocalls: A Survey of Techniques against Telephone Spam", In *Proceedings of 37th IEEE Symposium on Security and Privacy*, 2016.
- [TFP+06] H. Tschofenig, R. Falk, J. Peterson, J. Hodges, D. Sicker, and J. Polk. "Using SAML to Protect the Session Initiation Protocol (SIP). *Network Management of Global Internetnetworking*, Volume 20 Pages:14–17, 2006.
- [TS13] K. Toyoda and I. Sasase. "SPIT Callers Detection with Unsupervised Random Forests Classifier", In *Proceedings of IEEE ICC*, 2013.
- [UT08] M. Uemura and T. Tabata. "Design and Evaluation of a Bayesian-filter-based Image Spam Filtering Method", In *Proceedings of ISA 2008*, 2008.
- [VHS09] J. Vosecky, Dan Hong, and V.Y. Shen. "User Identification across Multiple Social Networks", In *Proceedings of Ist Networked Digital Technologies*, 2009.
- [WAB+09] Y.S. Wu, V. Apte, S. Bagchi, S. Garg, and N. Singh. "Intrusion Detection in Voice Over IP Environments", *International Journal of Information Security*, Volume 8 Pages:153–172, 2009.
- [WA10] A. Hai Wang. "Don't follow me: Spam detection in twitter", In *Proceedings of the 2010 SECRIPT International Conference on*, 2010.
- [WBS+09] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao. "User Interactions in Social Networks and their Implications", In *Proceedings of 4th ACM European conference on Computer systems*, 2009.
- [WBS+09] Y.S. Wu, S. Bagchi, N. Singh, and R. Wita. "Spam Detection in Voice-Over-IP Calls through Semi-Supervised Clustering", In *Proceedings of 39th Annual IEEE/IFIP DSN*, 2009.

- [WIP11] De. Wang, D. Irani, and C. Pu. "A Social-spam Detection Framework", In *Proceedings of 8th CEAS*, 2011.
- [WMH07] F. Wang, Y. Mo, and B. Huang. "P2P-AVS: P2P Based Cooperative VoIP Spam Filtering", In *Proceedings of IEEE WCNC*, 2007.
- [XFH15] Cao Xiao, David Mandell Freeman, and Theodore Hwa. "Detecting clusters of fake accounts in online social networks", In *Proceedings of the 8th AISec*, 2015.
- [XYY+13] J. Xue, Z. Yang, X. Yang, X. Wang, L. Chen, and Y. Dai. "Votetrust: Leveraging friend invitation graph to defend against social network sybils", In *Proceedings of 2013 INFOCOM*, 2013.
- [YL07] E. A. Yavuz and V. C. M. Leung. Modeling channel occupancy times for voice traffic in cellular networks. In *2007 IEEE ICC*, 2007.
- [YPC+11] T. Yao, S. Pin-Chieh, and C. Ming-Syan. "Cosdes: A Collaborative Spam Detection System with a Novel E-Mail Abstraction Scheme", *IEEE Transactions on Knowledge and Data Engineering*, Volume 23 Pages:669–682, 2011.
- [ZG09] R Zhang and A. Gurkov. "Collaborative Reputation-based Voice Spam Filtering", In *Proceedings of DEXA 09.*, 2009.
- [ZL09] R. Zafarani and H. Liu. "Connecting Corresponding Identities Across Communities", In *Proceedings of ICWSM 09*, 2009.
- [ZL13] R. Zafarani and H. Liu. "Connecting Users Across Social Media Sites: A Behavioral-modeling Approach", In *Proceedings of 19th ACM SIGKDD*, 2013.
- [ZLC+06] R. Zheng, J. Li, H. Chen, and Z. Huang. "A Framework for Authorship Identification of Online Messages: Writing-style Features and Classification Techniques", *Journal of the American Society for Information Science and Technology*, Volume 57 Pages:378–393, 2006.
- [ZWY+07] R. Zhang, X. Wang, X. Yang, and X. Jiang. "Billing Attacks on SIP-based VoIP Systems", In *Proceedings of Proceedings of 1st USENIX workshop on Offensive Technologies*, 2007.