

RE-IDENTIFICATION OF INDIVIDUALS IN GENOMIC DATA-SHARING BEACONS VIA ALLELE INFERENCE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Nora von Thenen
October 2017

Re-Identification of Individuals in Genomic Data-Sharing Beacons via
Allele Inference

By Nora von Thenen

October 2017

We certify that we have read this thesis and that in our opinion it is fully adequate,
in scope and in quality, as a thesis for the degree of Master of Science.

A. Ercument Cicek(Advisor)

Mehmet Koyuturk

Nurcan Tuncbag

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

RE-IDENTIFICATION OF INDIVIDUALS IN GENOMIC DATA-SHARING BEACONS VIA ALLELE INFERENCE

Nora von Thenen

M.S. in Computer Engineering

Advisor: A. Ercument Cicek

October 2017

Genomic datasets are often associated with sensitive phenotypes. Therefore, the leak of membership information is a major privacy risk. Genomic beacons aim to provide a secure, easy to implement, and standardized interface for data sharing by only allowing yes/no queries on the presence of specific alleles in the dataset. Previously deemed secure against re-identification attacks, beacons were shown to be vulnerable despite their stringent policy. Recent studies have demonstrated that it is possible to determine whether the victim is in the dataset, by repeatedly querying the beacon for his/her single nucleotide polymorphisms (SNPs). In this thesis, we propose a novel re-identification attack and show that the privacy risk is more serious than previously thought. Using the proposed attack, even if the victim systematically hides informative SNPs (i.e., SNPs with very low minor allele frequency -MAF-), it is possible to infer the alleles at positions of interest as well as the beacon query results with very high confidence. Our method is based on the fact that alleles at different loci are not necessarily independent. We use the linkage disequilibrium and a high-order Markov chain-based algorithm for the inference. We show that in a simulated beacon with 65 individuals from the CEU population, we can infer membership of individuals with 95% confidence with only 5 queries, even when SNPs with MAF less than 0.05 are hidden. This means, we need less than 0.5% of the number of queries that existing works require, to determine beacon membership under the same conditions. We further show that countermeasures such as hiding certain parts of the genome or setting a query budget for the user would fail to protect the privacy of the participants under our adversary model.

Keywords: beacon, genome privacy, re-identification attack.

ÖZET

GENOM VERİSİ PAYLAŞAN BEACON SİSTEMLERİNE KARŞI ALEL ÇIKARIMI YAPAN KİMLİK TESPİTİ ATAKLARI

Nora von Thenen

Bilgisayar Mühendisliği, Yüksek Lisans

Tez Danışmanı: A. Ercüment Çiçek

Ekim 2017

Genom veri setleri genellikle hassas fenotipler ile ilişkilidirler. Bu nedenle bir kişinin veri setinde olduğunun anlaşılması büyük bir mahremiyet riskidir. Beacon sistemleri veri paylaşımı için güvenli, kolay kurulabilir ve standardize bir arayüz sunmayı amaçlar. Bu sistemler sadece kendilerine sorulan, belli alellerin veri setinde olup olmadığına dair evet/hayır sorularını cevaplarlar. Bu kısıtlayıcı prosedür nedeniyle kimlik tespiti ataklarına karşı güvenilir oldukları düşünülen beacon sistemlerinin, risk taşıdığı gösterilmiştir. Yakın zamandaki çalışmalar, bir kişinin veri setinde olup olmadığını anlamanın, beacon sistemlerini bu kişinin nokta mutasyonları ile defalarca sorgulayarak mümkün olabileceğini göstermiştir. Bu tezde özgün bir kimlik tespiti saldırısı tanımlanmakta ve riskin önceden düşünüldüğünden daha büyük olduğu gösterilmektedir. Bu saldırı ile, saldırıya uğrayan kişinin tanımlayıcı mutasyonları gizlenmiş olsa bile, bu aleller çıkarım yolu ile bulunabilir ve beacon sisteminin verdiği cevaplar yüksek güven ile tahmin edilebilir. Algoritma, farklı pozisyonlardaki alellerin bağımsız olmamasını temel alarak çalıřır ve linkaj dengesizliği ile yüksek seviye Markov zinciri kullanmaktadır. 65 Avrupalı (CEU) bireyi içeren beacon sistemi simülasyonunda, sadece 5 sorgu ile bir kişinin veri setinde olup olmadığını %95 güvenilirlik ile belirleyebileceğimiz gösterilmiştir (minör alel frekansı 0.05'ten küçük olan olan mutasyonlar sistematik olarak gizlendiğinde bile). Bu rakam, diğer metotların gerek duyduğu sorgu sayısının %0.5'ine denk gelmektedir. Son olarak, literatürde önerilmiş olan, genom verisinin bazı bölgelerinin saklanması ya da kişi başına bir sorgu bütçesi atanması gibi savunma metotlarının da bizim modelimizde katılımcıların mahremiyetini korumakta yetersiz kaldığı gösterilmiştir.

Anahtar sözcükler: beacon, genom mahremiyeti, kimlik tespit saldırısı.

Acknowledgement

I would first like to thank my thesis advisor Asst. Prof. A. Ercument Cicek for giving me the opportunity to study under his supervision and for leading me through this learning process by providing valuable feedback whenever I needed it. Furthermore, I would like to express my gratitude to Asst. Prof. Erman Ayday for his support in this project and for always finding time to answer my questions.

I would also like to thank Anisa Halimi, Saharnaz E. Dilmaghani and Didem Demirağ for all the discussions and their support throughout my entire Master's program.

Finally, I would like to express my gratitude to my family in Germany and Turkey. Especially, I would like to thank my husband, who was always by my side, encouraged me and never failed to support me, our Turkish family for making me feel home and welcome, my sister for her unconditional support and love since I was little and my parents for never questioning my decisions but always trusted and believed in me. I would not be who I am and where I am today without each one of you. Thank you.

Contents

1	Introduction	1
2	Related Work	6
2.1	Shringarpure and Bustamante’s Attack	7
2.2	Optimal Attack	9
3	Query Inference Attack	11
4	Genome Inference Attack	16
5	Results	19
5.1	Experimental Set-Up	19
5.2	Re-identification on a simulated Beacon	22
5.3	Re-identification on Existing Beacons	23
6	Discussion	27

7 Conclusion	30
---------------------	-----------

A LRT - Power Calculation	34
----------------------------------	-----------



List of Figures

1.1	System Model of a Beacon Query	2
1.2	Single Nucleotide Polymorphism DNA variations that commonly occur within a population.	4
2.1	Models of the four attacker models (a) SB attack [1], (b) Optimal attack [2], (c) QI-attack and (d) GI-attack. Upper-case letters represent the major allele at a SNP position and the lower-case letters the corresponding minor allele. The SB attack randomly selects the minor allele from heterozygous SNP positions of the victim and queries those. The Optimal attack first sorts the heterozygous SNPs with respect to their MAF and queries for the minor alleles starting with the lowest frequency. Depending on the threshold t , SNPs with an $MAF < t$ are hidden and not available to the attacker. The QI-attack is identical to the Optimal attack but extends it by inferring beacon answers using correlations between SNP pairs. The GI-attack infers the hidden SNPs using a high-order Markov chain and queries the beacon for the minor alleles of those positions.	10

3.1	An example SNP network, containing of 5 nodes (i.e. SNPs). The SNP network is a directed graph, where the edges resemble the correlation. This example shows a completely connected graph, not all SNP subnetworks are completely connected.	13
3.2	Four attacker models: SB attack [1], Optimal attack [2], QI-attack, and GI-attack and their background knowledge for two scenarios are shown. In the first scenario $t = 0$ and in the second scenario $t > 0$, where t is the threshold up to which SNPs of the victim with an $\text{MAF} < t$ are hidden as a countermeasure. In Scenario 1, the attacker has access to the full genome of the victim (no hidden SNPs). In Scenario 2, SNPs with an $\text{MAF} < t$ are hidden and the attacker has partial access to the genome of the victim.	15
5.1	Experiments with a simulated beacon with 65 members (blue and red). 40 individuals who build the case set (orange) and are not in the beacon and 20 individuals (red) who build up the control set and are beacon members.	20
5.2	(a) Close-up of the power curves, where number of queries < 10 . (b) Power curves of the Optimal attack [2], the QI-attack, and the GI-attack for different thresholds of t on a beacon with 65 members constructed with individuals from the CEU dataset of the HapMap project. t indicates the threshold up to which SNPs with an $\text{MAF} < t$ are hidden as a countermeasure.	21
5.3	The GI-attack for $t = 0.03$ with the high-order Markov chain trained on the victim's population (CEU) in comparison to the high-order Markov chain trained on a different population (here MEX) from the HapMap dataset.	26

A.1	Example Λ distributions for 3 of the 40 case and 3 of the 20 control individuals of the experiments with a simulated beacon in Section 5.2 for the Optimal attack.	35
A.2	Example Λ distributions for 3 of the 40 case and 3 of the 20 control individuals of the experiments with a simulated beacon in Section 5.2 for the QI-attack.	35



List of Tables

3.1	Relationship between Linkage Disequilibrium (LD) measured by D between the SNPs A and B and their allele frequencies.	11
4.1	Comparison of different values for k (order of the high-order Markov chain). # of same markers shows how many markers that were inferred by the Markov chain were also asked in the Optimal attack. Distance to real response shows the amount of queries the inferred response differs from the Optimal attack's response (on average). # of people not inferred shows the amount of people that could not be inferred for that k	17
5.1	Number of queries needed to re-identify individuals for the SB attack [1], the Optimal attack [2], the QI-attack, and the GI-attack for different thresholds of t on a Beacon with 65 members constructed with individuals from the CEU dataset of the HapMap project. t indicates the threshold up to which SNPs with an MAF $< t$ are hidden. As the GI-attack concentrates on inferring hidden parts of the genome, we do not compute the case $t = 0$ (nothing is hidden) for the GI-attack.	22

5.2	Number of queries required to receive a “no” within 1000 queries to existing beacons using an individual from PGP [3] when $t = \{0, 0.03, 0.05\}$ for the Optimal attack [2], the QI-attack, and the GI-attack. Here, empty answers are not considered as a “no” response.	25
5.3	Number of queries required to receive a “no” within 1000 queries to existing beacons using an individual from PGP [3] when $t = \{0, 0.03, 0.05\}$ for the Optimal attack [2], the QI-attack, and the GI-attack. Here, empty answers are considered as a “no” response.	25

Chapter 1

Introduction

Exciting times are on the horizon for the genomics field with the announcement of the precision medicine initiative [4] which was followed by the \$55 million funding by NIH for the sequencing of a million individuals and AstraZeneca’s project of sequencing two million individuals [5]. Even though such million-sized genomic datasets are invaluable resources for research, sharing the data is a big challenge due to re-identification risk. Several studies in the last decade have shown that removal of personal identifiers from genomic data is not enough and that individuals can be re-identified using allele frequency information [6, 7, 8, 9, 10].

Genomic data-sharing beacons (referred to as beacons from now on) are the gateways that let users and data owners exchange information without -in theory- disclosing any personal information. A user who wants to apply for access to the dataset can learn whether individuals with specific alleles of interest are present in the beacon through an online interface. More specifically, the user submits a query, asking whether a genome exists in the beacon with a certain nucleotide at a certain position, and the beacon answers “yes” or “no” (Figure 1.1). Beacons are easy to set up systems that provide very restricted access to the stored data. The Beacon Project is an initiative by the Global Alliance for Genomics and

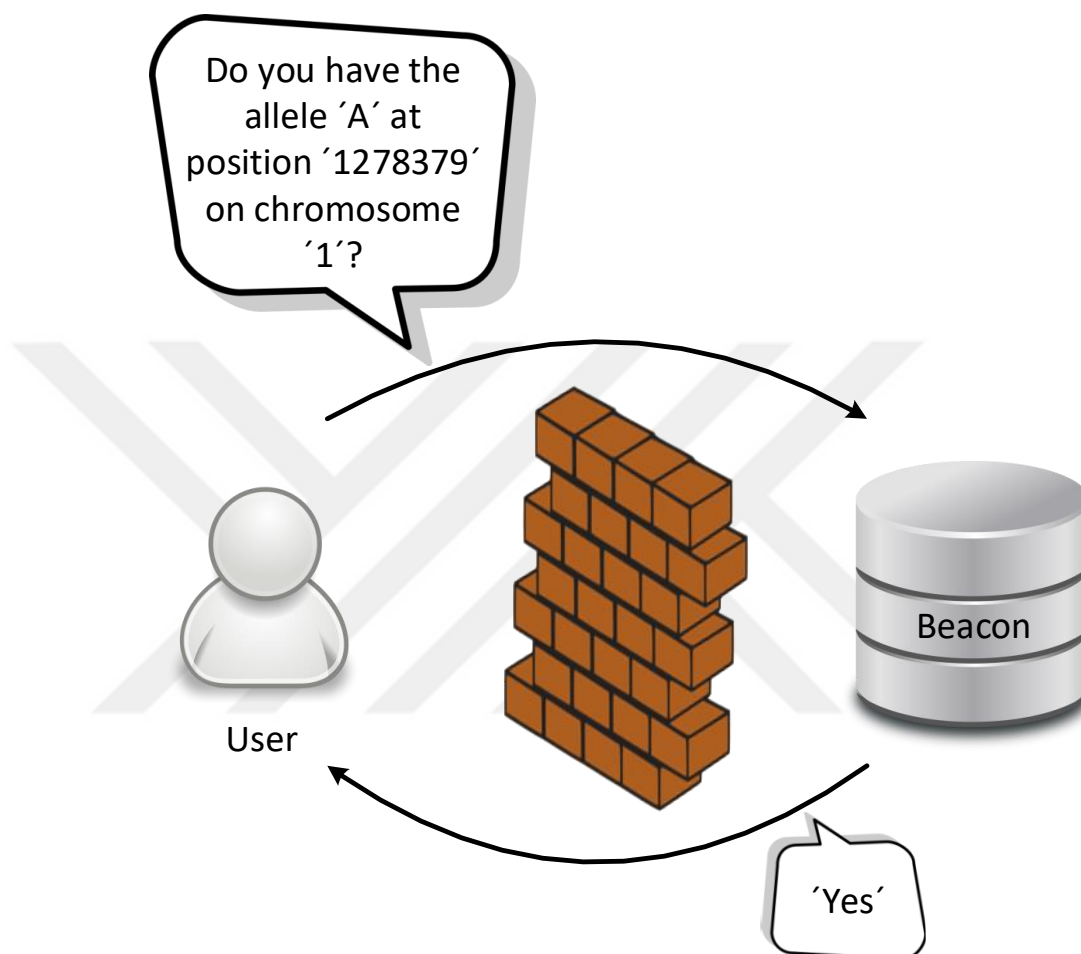


Figure 1.1: System Model of a Beacon Query

Health (GA4GH) which creates policies to ensure standardized and secure sharing of genomic data. Beacons were considered safe as allele frequencies are not involved in the query result and the binary answers for allele presence seem far from being informative for an attack. However, in 2015, Shringarpure and Bustamante introduced a likelihood-ratio test (LRT) that predicts if an individual is in the beacon or not, by repeatedly querying the beacon for single nucleotide polymorphisms¹ (SNPs, an example is shown in Figure 1.2) of the victim (dubbed

¹Single Nucleotide Polymorphisms are DNA variations that commonly occur within a population. If an individual has two different alleles at one SNP position, that position is referred to as a heterozygous position. One allele generally occurs more frequently and is therefore called the major allele, the less frequent allele is referred to as the minor allele.

the SB attack) [1]. The method does not use the allele frequencies and can compensate sequencing errors. They show that they could re-identify an individual in a beacon with 65 European individuals from the 1000 Genomes Project [11] with 250 queries (with 95% confidence). In their scheme, both the queries posed and the answers received from the beacon are assumed to be independent, therefore the hypothesis is tested based on a binomial test. Very recently, the work by Raisaro *et al.* showed that if the attacker has access to the MAFs of the population, s/he can sort the victim’s SNPs and query the SNPs starting from the one with the lowest MAF (dubbed the Optimal attack) [2]. SNPs are DNA variations that commonly occur within a population as shown in Figure 1.2. Unlike the SB attack, queries are not random in this case. As low MAF SNPs are more informative, Raisaro *et al.* show that fewer queries are needed to re-identify an individual. Furthermore, Raisaro *et al.* proposed countermeasures against re-identification attacks such as adding noise to the beacon results and assigning a budget to beacon members which limits the number of informative queries that can be asked on each member.

In this thesis, we introduce two new inference-based attacks that (i) carefully select the SNPs to be queried and predict query results of the beacon, and (ii) infer hidden or missing alleles of a victim’s genome. First, we show that if the queried locus is in linkage disequilibrium² (LD) with others, it is enough to query for that particular allele, as the attacker can infer the answers of the other alleles with high confidence [12]. We refer to this method as the QI-attack (query inference attack). Second, we introduce the GI-attack (genome inference attack) which recovers hidden parts of a victim’s genome by using a high-order Markov chain [13].

We show that in a simulated beacon with 65 European individuals (CEU) from the HapMap Project [14], our QI-attack requires 282 queries and our GI-attack requires only 5 queries on average to re-identify an individual, whereas the SB attack requires 19,525 queries and the Optimal attack requires 415 queries,

²Linkage disequilibrium (LD) is a measure to show how correlated two SNPs are. If two SNPs have a high LD value, they are likely to be inherited together [13]. Generally, SNPs that are close to each other on the DNA sequence are correlated, as the DNA is inherited in chunks rather than single positions. The LD value can be used to calculate the correlation of two specific nucleotides at two loci (SNP positions).

all at the 95% confidence level when the victim’s SNPs with MAFs < 0.03 are hidden. Therefore, the attacker models presented here can efficiently work when certain regions in the genome of the victim are systematically hidden as a security countermeasure. The number of queries required by the SB and the Optimal attacks substantially increase as more SNPs are concealed, while the GI-attack still requires only a few queries on average. Finally, we show that the QI-attack can still re-identify individuals despite the stringent query budget countermeasure proposed by [2] and the beacon censorship countermeasure proposed by [1].

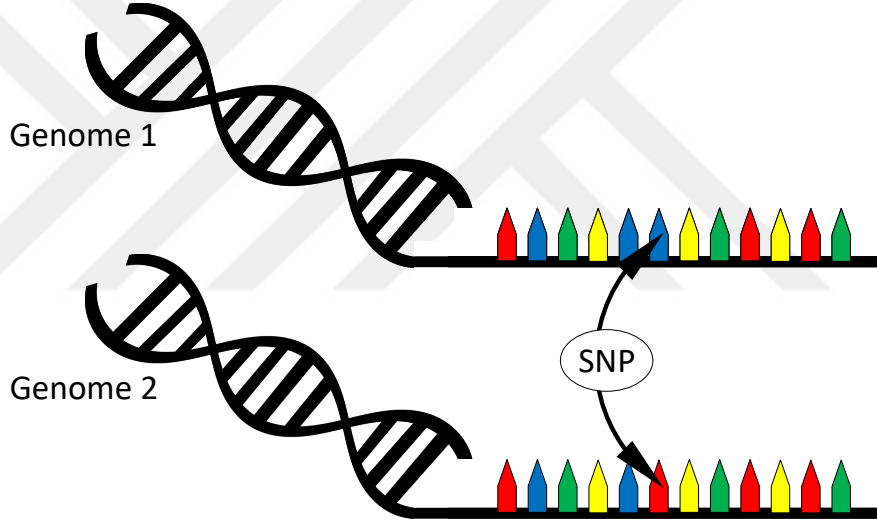


Figure 1.2: Single Nucleotide Polymorphism DNA variations that commonly occur within a population.

We demonstrate that the beacons are more vulnerable than previously thought and that the proposed countermeasures in the literature still fail to protect the privacy of the individuals. The contributions of this thesis can be summarized as follows:

- By inferring query results and alleles at certain positions, we show that it is possible to significantly decrease the number of required queries compared to other attacks in the literature [1, 2].
- We show that beacons are vulnerable even under a weaker adversary model, in which informative parts of a victim’s genome are concealed (such as all

SNPs with an MAF less than a threshold).

- We discuss the feasibility and the effectiveness of the proposed countermeasures in the literature and show that using the presented attack models, the participants are still under risk.

In this thesis, we will firstly give a literature review and present the related works of this field. Then, we will describe the two main algorithms that this thesis is based on in more detail (Chapter 2). In Chapter 3, we will present the methodology of the first proposed attack and in Chapter 4, we will show the methodology of the second proposed attack. Finally, we will present the results in Chapter 5 and discuss the results of this thesis in Chapter 6.

Chapter 2

Related Work

In this chapter, we will first give an overview of existing works in the field of genomic privacy and SNP correlations. In Section 2.1, we will introduce the attack proposed by Shringarpure and Bustamante in 2015 (referred to as SB attack) [1]. Raisaro *et al.* proposed an extended version of the SB attack [2], which we will introduce in Section 2.2 (referred to as Optimal attack).

By developing a statistical test in 2008, Homer *et al.* showed how the DNA of an individual can be identified within a complex genomic mixture, even if only 0.1% of the DNA in the mixture belongs to that individual. For their experiments, Homer *et al.* used high-density SNP genotyping microarray data [6]. In 2009, Jacobs *et al.* introduced a likelihood-ratio test that could determine whether an individual is part of a genome-wide association study (GWAS) and in which group, i.e. case or control that individual is. For their method, they only needed the genome of the individual and the genotype frequencies of each of the groups. Furthermore, previous works in the field of genomics and privacy have shown that it is possible to increase the success of genomic re-identification attacks by including linkage disequilibrium information of SNPs into the attacker model. Namely, Wang *et al.* showed in 2009 that individuals can be re-identified by using (i) publicly available SNP-to-disease correlation information, and (ii) SNPs in linkage disequilibrium (LD) [15]. To protect the genomic data used in medical

tests and personalized medicine Ayday *et al.* proposed a privacy-preserving disease susceptibility test (PDS) in 2013, which also includes LD information [16]. In 2013, Humbert *et al.* showed how LD can be used to build a framework to reconstruct the genomes of relatives from the genome of one family member [12]. To protect the privacy of individuals in genomic studies, Sankararaman *et al.* developed the tool SecureGenome that can determine SNPs in LD and outputs the SNPs that can be exposed without endangering the dataset members’ privacy [7]. Nevertheless, these works have not considered the power of high-order correlation within the genome (e.g., instead of pairwise correlations).

Using high-order correlations instead of pairwise LD correlations has already been studied by Gorelick *et al.* (2004) [17], Kim *et al.* (2008) [18] and Feng *et al.* (2008) [19]. In 2015, Samani *et al.* presented an inference attack that is based on high-order SNP correlations by using a high-order Markov chain [13].

As mentioned above, beacon servers are open to re-identification attacks and therefore put their members’ privacy at risk. The following subsections 2.1 and 2.2 explain the two latest proposed re-identification attacks in more detail.

2.1 Shringarpure and Bustamante’s Attack

In 2015, Shringarpure and Bustamante showed they can re-identify a person and reveal phenotype information with high accuracy by querying a beacon 250 times using real genotype data (SB attack). The likelihood-ratio test (LRT) proposed by Shringarpure and Bustamante is based on the “yes” responses of the queried beacon using a target’s VCF¹ file [1]. Their work uses the same attacking strategy as Homer *et al.* in 2008, which concentrates on heterozygous positions of the victim to be re-identified [6]. That is, querying only SNPs with two different alleles in a position by only considering bi-allelic SNP positions (e.g. “AT”). The queried SNPs are picked randomly from the victim’s heterozygous SNP positions. The null hypothesis (H_0) refers to the query genome being not in the beacon

¹Variant Call Format (VCF) is a file format to store the SNP data of an individual.

database. Under the alternative hypothesis (H_1) the query genome is a member of the beacon. Thus, an ideal response sequence to prove membership of the victim would be expected to consist of only positive responses of the beacon, such as: $x_1 = x_2 \dots = x_n = 1$, where x_i represents the response of the beacon to query i , n is the number of queries and 1 corresponds to a "yes" response. However, due to possible differences between a person's sequence in the beacon and the copy at hand, Shringarpure and Bustamante introduced δ , which represents the probability of such an error. The log-likelihood under the null hypothesis has been defined as shown in Equation 2.1.

$$L_{H_0}(R) = \sum_{i=1}^n x_i \log(1 - D_N) + (1 - x_i) \log(D_N), \quad (2.1)$$

where R is the response set and D_N the probability that no individual in the beacon has the queried allele at that position. x_i is the answer of the beacon to the query at position i (1 for yes, 0 for no), and n is the total number of posed queries. Accordingly, the log-likelihood of the alternative hypothesis has been stated as shown in Equation 2.2.

$$L_{H_1}(R) = \sum_{i=1}^n x_i \log(1 - \delta D_{N-1}) + (1 - x_i) \log(\delta D_{N-1}), \quad (2.2)$$

where D_{N-1} represents the probability of no individual except from the queried person having the same SNP. δ represents a possible sequencing error. By combining both hypotheses, Shringarpure and Bustamante define the log-likelihood ratio test (LRT) as shown in Equation 2.3.

$$\Lambda = L_{H_0}(R) - L_{H_1}(R). \quad (2.3)$$

The LRT statistic can be written as shown in Equation 2.4.

$$\Lambda = nB + C \sum_{i=1}^n x_i, \quad (2.4)$$

where B and C are defined as $B = \log(D_N/\delta D_{N-1})$ and $C = \log(\delta D_{N-1}(1 - D_N)/D_N(1 - \delta D_{N-1}))$, respectively. Figure 2.1(a) illustrates the model for this attack.

2.2 Optimal Attack

The Optimal attack introduced by Raisaro *et al.* (2016) [2] is an extension to the work of Shringarpure and Bustamante. It also integrates publicly available minor allele frequency (MAF) information into the attackers background knowledge. In this attack, the victim's SNPs are sorted with respect to their MAFs. The beacon is queried accordingly, starting from the first heterozygous SNP with the lowest MAF. The methodology is visualized in Figure 2.1(b). In this setting, the computations of D_{N-1} and D_N depend on the position i and change at each iteration as shown in Equations 2.5 and 2.6.

$$D_{N-1}^i = (1 - f_i)^{2N-2}, \quad (2.5)$$

$$D_N^i = (1 - f_i)^{2N}, \quad (2.6)$$

where f_i represents the MAF of the SNP at position i . Accordingly, Λ is determined by Equation 2.7.

$$\begin{aligned} \Lambda &= \sum_{i=1}^n \log\left(\frac{D_N^i}{\delta D_{N-1}^i}\right) + \log\left(\frac{\delta D_{N-1}^i (1 - D_N^i)}{D_N^i (1 - \delta D_{N-1}^i)}\right) x_i \\ &= \sum_{i=1}^n \log(\delta^{-1} (1 - f_i)^2) + \log\left(\frac{\delta}{(1 - f_i)^2} \frac{1 - (1 - f_i)^{2N}}{1 - \delta (1 - f_i)^{2N-2}}\right) x_i \end{aligned} \quad (2.7)$$

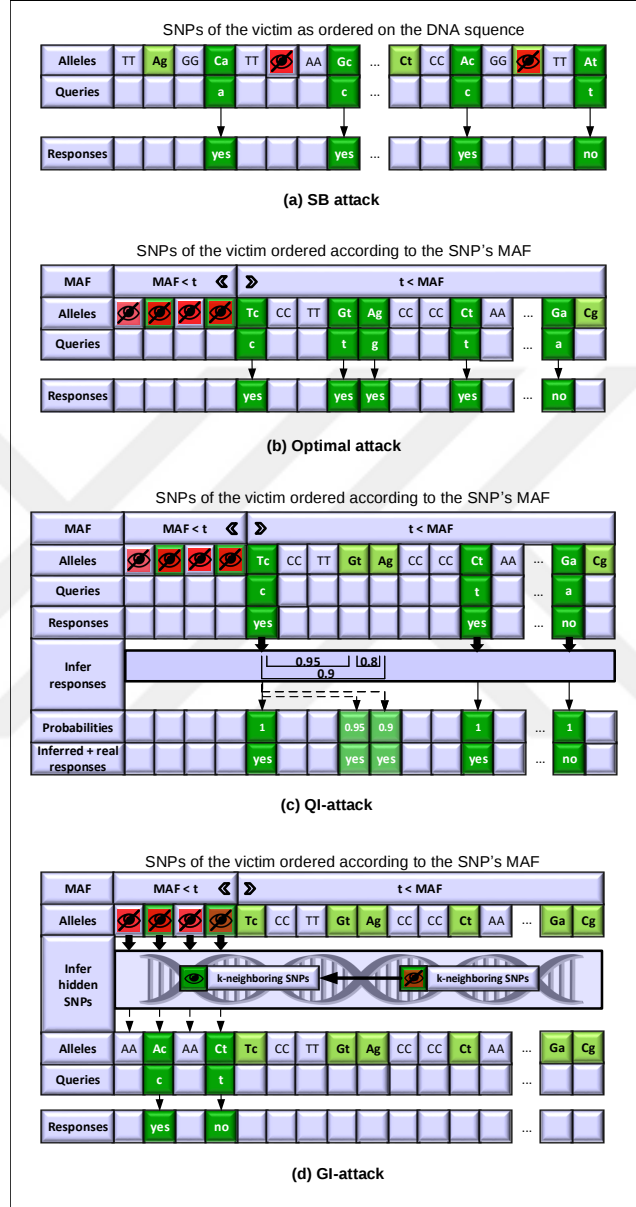


Figure 2.1: Models of the four attacker models (a) SB attack [1], (b) **Optimal attack** [2], (c) **QI-attack** and (d) **GI-attack**. Upper-case letters represent the major allele at a SNP position and the lower-case letters the corresponding minor allele. The SB attack randomly selects the minor allele from heterozygous SNP positions of the victim and queries those. The Optimal attack first sorts the heterozygous SNPs with respect to their MAF and queries for the minor alleles starting with the lowest frequency. Depending on the threshold t , SNPs with an $MAF < t$ are hidden and not available to the attacker. The QI-attack is identical to the Optimal attack but extends it by inferring beacon answers using correlations between SNP pairs. The GI-attack infers the hidden SNPs using a high-order Markov chain and queries the beacon for the minor alleles of those positions.

Chapter 3

Query Inference Attack

For the QI-attack, the attacker incorporates pairwise SNP correlations in order to infer beacon responses. That is, beacon responses for SNPs which are correlated to a queried SNP, are inferred. The attacker uses the LD value of a SNP pair to calculate the correlation of two minor alleles at two loci. The probability of two minor alleles at two loci occurring together can be calculated as p_2q_2 , where p_2 is the minor allele frequency of SNP A (with minor allele a) and q_2 is the minor allele frequency of SNP B (with minor allele b). This probability can increase or decrease if SNP A and SNP B are in LD. If their LD value increases the likelihood of the two minor alleles occurring together the term D is added to the probability. This can be calculated as follows; $Pr(ab) = p_2q_2 + D$ (as shown in Table 3.1), where D resembles the strength of the correlation of the two SNPs and is determined as $D = \sqrt{r^2(q_1q_2p_1p_2)}$.

Table 3.1: Relationship between Linkage Disequilibrium (LD) measured by D between the SNPs A and B and their allele frequencies.

	$Pr(A) = p_1$	$Pr(a) = p_2$
$Pr(B) = q_1$	$p_1q_1 + D$	$p_2q_1 - D$
$Pr(b) = q_2$	$p_1q_2 - D$	$p_2q_2 + D$

Algorithm 1: Stepwise procedure of the QI-attack, where AFs are the allele frequencies, S is the set of candidate SNPs to be queried and n the number of queries.

Require: VCF file, AFs, LD scores
 Read victim's VCF file
 Identify heterozygous SNP positions S of the victim
 Sort S based on MAFs (ascending order)
for all SNP_i in S **do**
 if SNP_i is member of SNP network **then**
 Get neighbors of SNP_i in SNP network
 Get cluster representative
 Query Beacon for cluster representative
 Infer response for neighboring SNPs
 else
 Query Beacon for SNP_i
 end if
end for

On this basis, the attacker constructs a SNP network that uses weighted, directed edges between SNPs in high LD (see Figure 3.1). The weight corresponds to the probability of the two minor alleles that are in LD occurring together. The probability of two minor alleles of two loci that are not in LD occurring together is equal to p_2q_2 , where p_2 is the minor allele frequency of SNP A (with minor allele a) and q_2 is the minor allele frequency of SNP B (with minor allele b). If A and B are in an LD relationship, the LD score between them increases the probability of two major or two minor alleles in these loci occurring together. As shown in Table 3 this leads to the formula: $Pr(ab) = p_2q_2 + D$, where D resembles the LD i.e. the strength of the correlation of the two SNPs. D is calculated as follows: $D = \sqrt{r^2(q_1q_2p_1p_2)}$, where q_1 and p_1 are the major allele frequencies and r^2 is a measure of LD. Furthermore, D' is needed, since it determines whether D is subtracted or added to the probability of two minor allele occurring together. $D' > 0.5$ implies D is added, whereas $D' < 0.5$ leads to a subtraction of D . In order to ensure high correlation between the SNPs, only LD relationships between SNP pairs with an r^2 value of more than 0.7 were considered.

Figure 2.1(c) illustrates the model for this attack. First, the attacker selects the SNPs to be queried. This step is identical to the Optimal attack and leads to

a set of candidate SNPs S to be queried, starting from the lowest MAF SNP_i . As a second step, the attacker determines the neighbors of each SNP to be queried in the SNP network. If the neighboring SNPs of SNP_i are in the SNP network and belong to the selected SNPs in S , the attacker can directly infer the query answers the beacon would have returned without posing a query for the neighboring SNPs. The steps of the attack are shown in Algorithm 1.

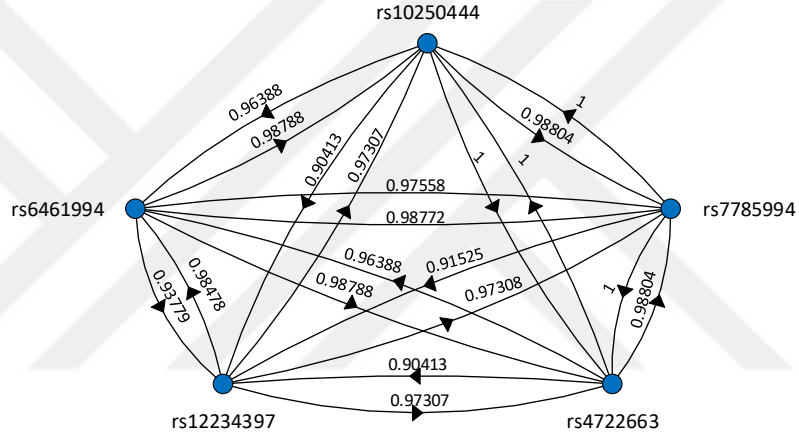


Figure 3.1: An example SNP network, containing of 5 nodes (i.e. SNPs). The SNP network is a directed graph, where the edges resemble the correlation. This example shows a completely connected graph, not all SNP subnetworks are completely connected.

The null hypothesis is then given as in Equation 3.1.

$$L_{H_0}(R) = \sum_{i=1}^n \left(x_i \log(1 - D_N^i) + (1 - x_i) \log(D_N^i) \right) + \sum_{j=1}^m \gamma x_i \log(1 - D_N^j) + \gamma (1 - x_i) \log(D_N^j) \quad (3.1)$$

where, n is the number of posed queries, m is the number of neighbors that can be inferred for each posed query x_i and γ corresponds to the confidence of the inferred answer, obtained from the SNP network.

$$\begin{aligned}
L_{H_1}(R) = & \sum_{i=1}^n \left(x_i \log(1 - \delta D_{N-1}^i) + (1 - x_i) \log(\delta D_{N-1}^i) \right. \\
& \left. + \sum_{j=1}^m \gamma x_i \log(1 - \delta D_{N-1}^j) + \gamma(1 - x_i) \log(\delta D_{N-1}^j) \right)
\end{aligned} \tag{3.2}$$

Accordingly, Equation 3.2 shows the alternative hypothesis. Λ is then calculated as shown in Equation 3.3.

$$\begin{aligned}
\Lambda = & L_{H_0}(R) - L_{H_1}(R) \\
= & \sum_{i=1}^n \left(x_i \log(1 - D_N^i) + (1 - x_i) \log(D_N^i) \right. \\
& \left. + \sum_{j=1}^m \gamma x_i \log(1 - D_N^j) + \gamma(1 - x_i) \log(D_N^j) \right) \\
& - \left[\sum_{i=1}^n \left(x_i \log(1 - \delta D_{N-1}^i) + (1 - x_i) \log(\delta D_{N-1}^i) \right) \right. \\
& \left. + \sum_{j=1}^m \gamma x_i \log(1 - \delta D_{N-1}^j) + \gamma(1 - x_i) \log(\delta D_{N-1}^j) \right] \\
= & \sum_{i=1}^n \left(x_i \log \left(\frac{1 - D_N^i}{1 - \delta D_{N-1}^i} \right) + (1 - x_i) \log \left(\frac{D_N^i}{\delta D_{N-1}^i} \right) \right. \\
& \left. + \sum_{j=1}^m \gamma x_i \log \left(\frac{1 - D_N^j}{1 - \delta D_{N-1}^j} \right) + \gamma(1 - x_i) \log \left(\frac{D_N^j}{\delta D_{N-1}^j} \right) \right) \\
= & \sum_{i=1}^n \left(\log \left(\frac{D_N^i}{\delta D_{N-1}^i} \right) + \log \left(\frac{\delta D_{N-1}^i (1 - D_N^i)}{D_N^i (1 - \delta D_{N-1}^i)} \right) x_i \right. \\
& \left. + \sum_{j=1}^m \log \left(\frac{D_N^j}{\delta D_{N-1}^j} \right) + \log \left(\frac{\delta D_{N-1}^j (1 - D_N^j)}{D_N^j (1 - \delta D_{N-1}^j)} \right) \gamma x_i \right)
\end{aligned} \tag{3.3}$$

By eliminating unnecessary queries, this attacker model can require less queries to the server than the Optimal attack by achieving the same response set.

Scenario Adversary + Background		Scenario 1 $t = 0$	Scenario 2 $t > 0$
 SB attack	Population of victim	+ AA AT AA CT	+  AT AA
 Optimal attack	Population of victim, corresponding MAF	+ GG CC TT	+  GG  
 QI - attack	Population of victim, corresponding MAF & LD	+ GT CC	+   CC
 GI - attack	Population of victim, corresponding MAF, High-Order Correlation		+ CC TA TT

Figure 3.2: Four attacker models: SB attack [1], Optimal attack [2], QI-attack, and GI-attack and their background knowledge for two scenarios are shown. In the first scenario $t = 0$ and in the second scenario $t > 0$, where t is the threshold up to which SNPs of the victim with an $\text{MAF} < t$ are hidden as a countermeasure. In Scenario 1, the attacker has access to the full genome of the victim (no hidden SNPs). In Scenario 2, SNPs with an $\text{MAF} < t$ are hidden and the attacker has partial access to the genome of the victim.

Chapter 4

Genome Inference Attack

Individuals may publicly share their genomes by taking necessary precautions, such as hiding their sensitive SNP positions with $MAFs < t$ (i.e. Scenario 2 in Figure 3.2). The GI-attack performs allele inference to recover hidden SNP positions and infers alleles at the victim's hidden loci. Note that, Scenario 1 is not applicable to the GI-attack, since in that scenario, the attacker can access SNPs with low MAFs. The attacker uses a high-order Markov chain to model SNP correlations as described by Samani *et al.* [13]. The SNPs are represented as 0, 1, or 2 depending on the number of minor alleles at the specific position of the genome. That is, major homozygous, heterozygous, and minor homozygous, respectively.

Figure 2.1(d) illustrates the model of this attack. Threshold t determines up until which value SNPs on the victim's genome are hidden. The attacker then infers SNP positions with $MAF < t$ that are not available in the victim's VCF file. Based on the victim's genome sequence, the attacker calculates the likelihood of the victim having a heterozygous position at the chosen position SNP_i as shown in Equation 4.1.

$$P_k(SNP_i) = P(SNP_i | SNP_{i-1}, SNP_{i-2}, \dots, SNP_{i-k}), \quad (4.1)$$

where k is the order of the Markov chain. In order to use a high-order Markov chain to infer hidden SNPs, genome sequences from public sources such as the

Table 4.1: Comparison of different values for k (order of the high-order Markov chain). # of same markers shows how many markers that were inferred by the Markov chain were also asked in the Optimal attack. Distance to real response shows the amount of queries the inferred response differs from the Optimal attack’s response (on average). # of people not inferred shows the amount of people that could not be inferred for that k .

k	# of same markers	distance to real response	# of people not inferred
3	13	0.6	2
4	15	0.62	1
5	14	0.79	5

1000 Genomes project or HapMap can be used to train the model. Accordingly, Samani *et al.* define the k^{th} -order model as shown in Equation 4.2.

$$P_k(SNP_i) = \begin{cases} 0 & \text{if } F(SNP_{i-k,i-1}) = 0 \\ \frac{F(SNP_{i-k,i})}{F(SNP_{i-k,i-1})} & \text{if } F(SNP_{i-k,i-1}) > 0 \end{cases}, \quad (4.2)$$

where $F(SNP_{i,j})$ is the frequency of occurrence of the sequence that contains SNP_i to SNP_j . The SNPs are ordered according to their physical positions on the genome. The model compares the SNPs in $SNP_{i,j}$ which are prior to SNP_i on the genome sequence, to the same SNP positions in the training dataset. If the training set contains other genomes with the same SNP sequence and these sequences are followed by a heterozygous position, we can calculate the probability of SNP_i being heterozygous for our victim. As an example, the victim’s 4th-order SNP sequence is [AA, AT, CC, TT]. We would now like to determine whether the following SNP_i , that is hidden in the VCF file at hand, is likely to be a heterozygous position. We identify other genomes in the training dataset with the same sequence and compute the frequency of this sequence being followed by a heterozygous position. That is, [AA, AT, CC, TT] $\rightarrow$ [AG]. As a result, we can determine the probability of the four SNPs being followed by a heterozygous position, which we can use to query the beacon.

If the calculated likelihood of the victim having a heterozygous position is high enough (in this case equal to 1), the attacker queries the beacon for the inferred SNP position, starting from the SNP with the lowest MAF. Algorithm 2 gives

a stepwise overview of the attack. The value of k is determined empirically as explained in Section 5.1.

Algorithm 2: Stepwise procedure of the GI-attack, where AFs are the allele frequencies, S is the set of candidate SNPs to be queried and n the number of queries

Require: VCF file, AFs, anonymized publicly available VCF files from the same population
Read victim's VCF file
Identify heterozygous SNP positions S of the victim
Sort S based on MAFs (ascending order)
Identify hidden SNP positions S'
for all SNP_i in S' **do**
 if SNP_i can be inferred **then**
 Query Beacon for SNP_i
 end if
end for

Chapter 5

Results

5.1 Experimental Set-Up

In this chapter, we first show our results for the four attacker models with data from the HapMap project [14] on a simulated beacon and on 9 existing beacons from the beacon-network API¹ (namely: Known VARiants, Broad Institute, 1000 Genomes Project, Cafe CardioKit, Wellcome Trust Sanger Institute, NCBI, ICGC, AMPLab, 1000 Genomes Project phase 3) using a person from the personal genomes project (PGP) [3]. By only giving presence information about alleles at certain positions of the genome, beacons seemed to be a safe way to share sensitive genomic information.

For our attacks, we consider two scenarios as shown in Figure 3.2. In the first scenario, the attacker has access to the full² genome of the victim as shown in Scenario 1. In Scenario 2, the attacker only has limited access to the genome, as the victim has systematically hidden sensitive SNP positions as a countermeasure.

In Section 5.2, we evaluate the performance of the four attacks on a beacon

¹<http://www.beacon-network.org>

²In this case “full” means that a part of the DNA of the victim (e.g. one chromosome) is available without systematically hidden SNP positions with low MAFs.

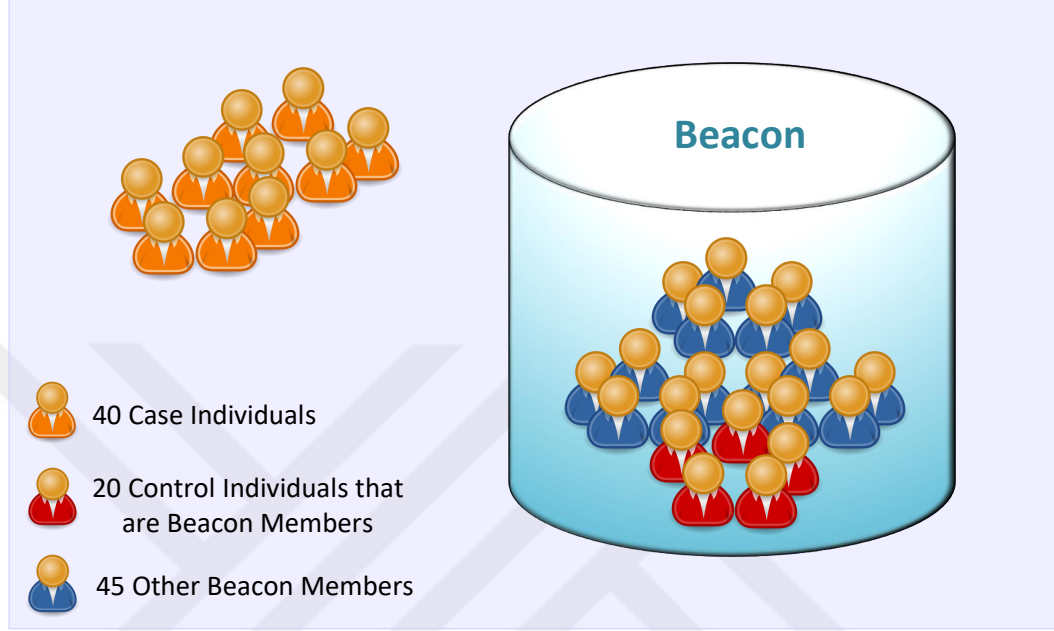


Figure 5.1: Experiments with a simulated beacon with 65 members (blue and red). 40 individuals who build the case set (orange) and are not in the beacon and 20 individuals (red) who build up the control set and are beacon members.

with 65 people from the Utah Residents with Northern and Western European Ancestry (CEU) population of the HapMap dataset and 40 additional people of the same population also from the HapMap dataset [14]. In order to test the beacon for both hypotheses, we used 60 individuals to test, of which 20 were members of the beacon (control set) and 40 were not in the beacon (case set) as illustrated in Figure 5.1. We decided to use the CEU population, because previous works (SB attack [1] and Optimal attack [2]) have also been evaluated on this population. The LD values, allele frequencies, and genotype data were obtained from the CEU dataset of the HapMap project [14].

For the GI-attack, we use a 4^{th} -order Markov chain. We chose $k = 4$ empirically, as it depends on the dataset that is used to train the model. As shown in Table 4.1, we considered (i) the number of markers that were inferred by the GI-attack and also asked by the Optimal attack, (ii) the euclidean distance between the number of queries needed by the Optimal attack and the GI-attack for all tested individuals and (iii) the number of people whose SNPs could not be

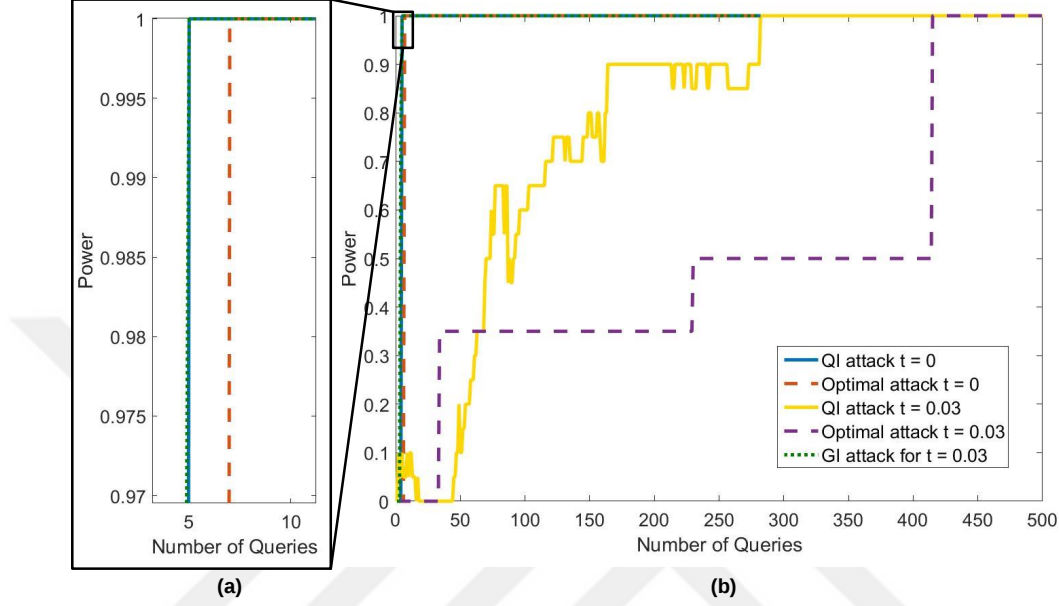


Figure 5.2: **(a)** Close-up of the power curves, where number of queries < 10 . **(b)** Power curves of the Optimal attack [2], the QI-attack, and the GI-attack for different thresholds of t on a beacon with 65 members constructed with individuals from the CEU dataset of the HapMap project. t indicates the threshold up to which SNPs with an $\text{MAF} < t$ are hidden as a countermeasure.

inferred due to too much missing data. We tested values for $k = 3$ to 5 to prevent over-fitting of the model, as our training set consists of only 100 individuals from the CEU population of the HapMap [14] dataset. Accordingly, we determined $k = 4$ as the best performing Markov chain for our dataset.

In order to test our methods on existing beacons in Section 5.3, we used the beacon-network API³ operated by GA4GH Beacon Network and one individual from PGP (Person’s id: PGP180/hu2D53F2) [3]. The beacons can return an empty response, that is, the beacon has no information at that position, a “no”-response and a “yes”-response. We consider two cases for the evaluation of the query results. In the first case, an empty answer is treated as a “no” (results shown in Table 5.2(2)), in the second case an empty answer is not treated as a “no”, as it is also possible that the beacon has a different copy of the victim’s

³<http://www.beacon-network.org>

Table 5.1: Number of queries needed to re-identify individuals for the SB attack [1], the Optimal attack [2], the QI-attack, and the GI-attack for different thresholds of t on a Beacon with 65 members constructed with individuals from the CEU dataset of the HapMap project. t indicates the threshold up to which SNPs with an $\text{MAF} < t$ are hidden. As the GI-attack concentrates on inferring hidden parts of the genome, we do not compute the case $t = 0$ (nothing is hidden) for the GI-attack.

	# of queries			
t	SB attack	Optimal attack	QI-attack	GI-attack
0	1,418	3	3	NA
0.03	19,525	270	160	2
0.05	56,759	1,495	1,031	2

genome (results shown in Table 5.3(3)).

5.2 Re-identification on a simulated Beacon

The power curves for the Optimal, the QI-attack and the GI-attack are calculated with false positive rate $\alpha = 0.05$ as shown in Figure 5.2 and the number of queries needed to receive the first negative response are shown in Table 5.1. We empirically build the null hypothesis using the 40 people who are not in the beacon and calculate Λ as stated for the different attacks in the corresponding sections. We reject the null hypothesis when $\Lambda < t_\alpha$. Similar to Raisaro *et al.* [2], we determine t_α from the null hypothesis with $\alpha = 0.05$. t_α is recalculated at each query. The power $1 - \beta$ is then the proportion of $\Lambda < t_\alpha$ for all individuals in the control set.

The performance of the attacks is significantly affected by threshold t of hidden SNPs. As t is increased only more common SNPs are available to the attacker which means that the likelihood of another individual in the beacon having the same allele increases. We queried the simulated beacon for each of the 40 individuals in the case set, where the SB attack was not able to receive a “no”

response within 100,000 queries, (i) for 4 people when SNPs with an $MAF < 0.04$ were hidden and (ii) for 12 people when SNPs with an $MAF < 0.05$ were hidden. Therefore, it was not possible to build the null hypothesis and reach 100% power. The Optimal and the QI-attack require a significantly higher number of queries to build the null hypothesis for increasing t . The GI-attack successfully determined the correct status for all 40 individuals despite the high threshold of t with only a few queries.

The SB attack requires the highest number of queries (1,400 - 56,800). The QI-attack requires 30% less number of queries on average compared to the Optimal attack. The GI-attack requires only 5 queries for all tested thresholds of t .

Compared to the monotonically increasing behavior of the power curves for the Optimal attack, the power curve for the QI-attack shows a zig-zag behavior.

The reason for the zig-zag behavior of the power is the difference in the number of inferred queries per posed query for those individuals without a “no” response. Note that, Λ decreases significantly for a large number of inferred queries. As t_α is determined empirically, t_α can also decrease which let’s the power drop. Nevertheless, when the null hypothesis is built and all necessary case individuals received a sufficient amount of “no” responses, the value of t_α stabilizes and the power reaches 100%. Please see Figure A.1 and Figure A.2 in Appendix A for example distributions of Λ under both hypotheses.

5.3 Re-identification on Existing Beacons

We selected an individual from the Personal Genomes Project (PGP) (Person’s id: PGP180/hu2D53F2) [3] as the victim. For the QI-attack, we used the same SNP network as for the simulated beacon in Section 5.2 that is based on the CEU population of HapMap. To determine if the person is a member of the beacons, we applied the SB attack as ground truth. Therefore, the null hypothesis (the individual is not in the beacon) is rejected if $p \text{ value} < 0.05$. The p value is

calculated as $P(x \geq k; x \text{ binomial}(n, 1 - D_N))$. The tested individual had a p value of 1 and is therefore not a member of any of the beacons. Furthermore, the meta data of the Kaviar beacon does not show our individual as a member.

Our experiments on existing beacons are shown in Table 5.2, where empty responses are ignored and in Table 5.3, where empty answers are considered as “no” responses. For 6 of the 9 tested beacons, we were able to determine that the victim is not a member of the beacons. For the Known VARiants (Kaviar), the Cafe CardioKit and the NCBI, it was not possible within 1,000 queries (Table 5.2). For the second case in Table 5.3, we could not detect the correct membership status for only 1 of the 9 beacons. As the large Kaviar beacon contains over 70,000 individuals, we only received “yes” responses within 1,000 queries. Overall, we observed that the experiments on real beacon support our findings from Section 5.2. That is, the Optimal and the QI-attack need more queries as t increases, the GI-attack is stable over all thresholds and the QI-attack requires less queries than the Optimal attack.

Unlike all other beacons, the 1000 Genome Project beacon requires less queries for re-identification as t is increased. One possible explanation is that the victim’s SNPs are being sorted based on the CEU population’s allele frequencies and SNPs that we query are not necessarily the rarest in the queried beacon. Furthermore, The SNP Network used here is also based on the CEU population and therefore, does not include all SNPs of the victim’s genome.

The GI-attack performed as expected, that is constant over the two tested thresholds of t and outperformed the Optimal attack [2] as well as the QI-attack for $t > 0$. For the 1000 Genomes Beacon the GI-attack requires the same amount of queries as the other attacks, as the number of queries needed are already very low.

In order to analyze the robustness of the GI-attack, we used a different training dataset to train the high-order Markov chain. The case and control individuals are the same as in the results shown above, that is from the CEU population. The high-order Markov chain was trained on the 77 individuals from the HapMap

Table 5.2: Number of queries required to receive a “no” within 1000 queries to existing beacons using an individual from PGP [3] when $t = \{0, 0.03, 0.05\}$ for the Optimal attack [2], the QI-attack, and the GI-attack. Here, empty answers are not considered as a “no” response.

Beacon Name t	Optimal attack 0 0.03 0.05	QI-attack 0 0.03 0.05	GI-attack 0.03 0.05
Known VARiants	— — —	— — —	— —
Broad Institute	2 2 2	2 2 2	1 1
1000 Genomes Project	4 3 2	4 3 2	3 3
Cafe CardioKit	— — —	— — —	— —
Wellcome Trust			
Sanger Institute	1 1 1	1 1 1	1 1
NCBI	— — —	— — —	— —
ICGC	1 — —	1 — —	1 1
AMPLab	20 45 73	20 40 73	39 39
1000 Genomes Project phase 3	20 130 250	20 116 250	48 48

Table 5.3: Number of queries required to receive a “no” within 1000 queries to existing beacons using an individual from PGP [3] when $t = \{0, 0.03, 0.05\}$ for the Optimal attack [2], the QI-attack, and the GI-attack. Here, empty answers are considered as a “no” response.

Beacon Name t	Optimal attack 0 0.03 0.05	QI-attack 0 0.03 0.05	GI-attack 0.03 0.05
Known VARiants	— — —	— — —	— —
Broad Institute	1 1 1	1 1 1	1 1
1000 Genomes Project	4 3 2	4 3 2	3 3
Cafe CardioKit	1 1 1	1 1 1	1 1
Wellcome Trust			
Sanger Institute	1 1 1	1 1 1	1 1
NCBI	1 1 1	1 1 1	1 1
ICGC	1 1 1	1 1 1	1 1
AMPLab	20 45 73	20 40 73	39 39
1000 Genomes Project phase 3	20 130 250	20 116 250	48 48

dataset “Mexican ancestry in Los Angeles” (MEX) and $k = 4$.

The GI-attack required 4 more queries to reach 100% power compared to the case when the correct population is used for training. However the power curves are similar as can be seen in Figure 5.3.

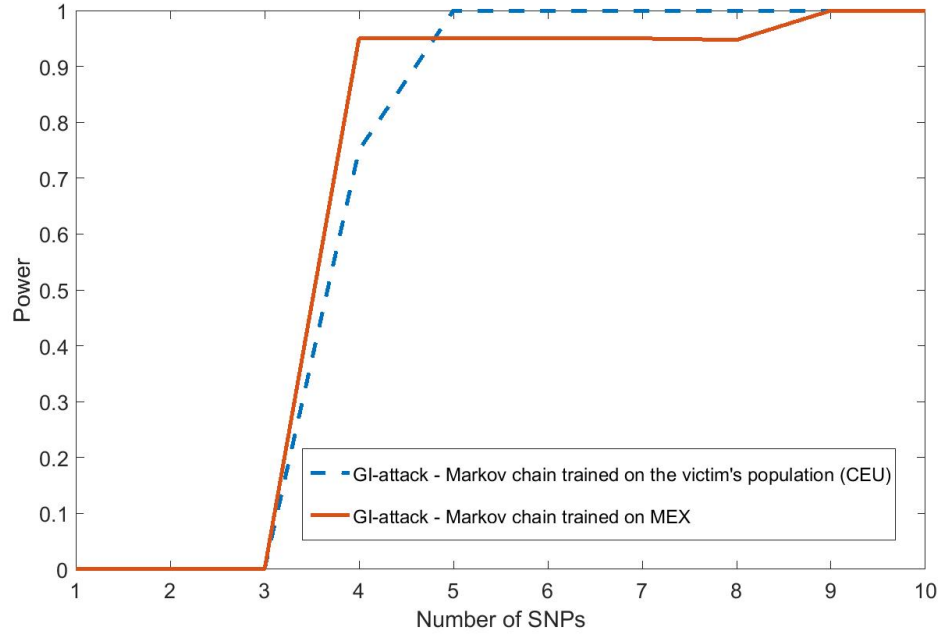


Figure 5.3: The GI-attack for $t = 0.03$ with the high-order Markov chain trained on the victim’s population (CEU) in comparison to the high-order Markov chain trained on a different population (here MEX) from the HapMap dataset.

Chapter 6

Discussion

Recent works by Shringarpure and Bustamante [1] and Raisaro *et al.* [2] have shown in 2015 and 2016, respectively that beacon servers fail to protect their members' privacy as successfully as previously thought. As beacons are often associated with a certain phenotype, the membership identification of an individual could leak sensitive phenotype information. Therefore, they proposed countermeasures to better protect the beacon members privacy.

In this thesis, we have shown that a small number of queries suffices to detect beacon membership with high confidence, even if data owners and the members of data-sharing platforms apply countermeasures to improve the security standards of the underlying datasets. By including publicly available information such as MAF, LD, and anonymized VCF files (from e.g. HapMap [14] or 1000 Genomes Project [11]) into the attacker model, these countermeasures can be overcome. Therefore, existing countermeasures fail to protect the sensitive genomic datasets.

Since the density and the size of the SNP network determines how many answers can be inferred with one query, the success of our QI-attack significantly depends on the structure of the underlying SNP network. The larger and denser the network becomes, the more query responses can be inferred. To reduce the

inference error of the query inference that the QI-attack performs, we only included SNP pairs that are in strong LD (i.e. $r^2 > 0.7$). Lowering this threshold would lead to more edges in the SNP network, and therefore to more inferred queries, yet at the same time increasing the inference error.

The GI-attack shows that even if genomes do not contain any SNPs with low MAFs, the individual’s privacy is not ensured, as it is possible to infer these positions from publicly available datasets (e.g. HapMap [14] or 1000 Genomes Project [11]). As shown in Section 5.3, the GI-attack still performs as good, even when the attacker trains the high-order Markov chain on a different population than the victim’s.

As it can be seen in Section 5.2, in Table 5.2 and in Table 5.3, our experiments on (i) simulated and (ii) existing beacons show that as t increases the SB attack [1], the Optimal attack [2] and the QI-attack require more queries to detect beacon membership, where the GI-attack is stable over all tested thresholds of t . Overall, our attacks require less queries than existing attacks (SB attack [1] and Optimal attack [2]). Additionally, Table 5.2 shows that for existing beacons the number of queries needed increases as t increases (Table 5.1).

Shringarpure and Bustamante, 2015 discussed different countermeasures, such as (i) increasing the beacon size, (ii) sharing only small genomic regions, (iii) using single population beacons, (iv) not publishing the metadata of a beacon, and (v) adding control samples to the beacon dataset [1]. Lately, Aziz *et al.*, 2017 proposed two algorithms which are based on randomizing the response set of the beacons with the goal of protecting beacon members’ privacy while maintaining the efficacy of the beacon servers [20]. Raisaro *et al.*, 2016 have analyzed the behavior of the beacon when applying three different countermeasures [2]. First, they propose the beacon should only respond “yes” for an allele it contains more than one time. That is, the allele in that position occurs at least two or three times. The second countermeasure adds noise to the responses of the beacon and therefore answers “no” instead of “yes” to some queries. However, this countermeasure significantly reduces the utility of the dataset and is unacceptable for researchers working on beacons. Instead, the beacon could return an empty

answer. Lastly, Raisaro *et al.* proposed a query budget as a countermeasure. That is, every member of the beacon is assigned with a certain budget that is reduced if a query to the beacon matches one of their SNP positions. As an example, if a user queries the beacon for allele A in position 1000 of chromosome 21, then the budget of every member with an allele A in that position is reduced. The amount of the budget reduction is determined based on the risk of the query, where the lower the allele frequency of the queried allele is, the higher the risk becomes. The budget is calculated as $b_i = \log(p)$, where Raisaro *et al.* use $p = 0.05$. The risk then is calculated as $r_i = -\log(1 - D_N^i)$. If the budget of a beacon member is depleted, the beacon stops including the member into the beacon responses.

An attacker using the QI-attack can use SNP correlations to overcome this budget countermeasure. Assuming the attacker is trying to identify beacon membership of the victim, by using the Optimal attack, a beacon applying the budget countermeasure would start returning false responses, which would lead to a wrong conclusion by the attacker. For instance, in our simulated beacon as described in Section 5.3, an attacker needs 7 queries to determine beacon membership of the victim (individual “NA12272” of the HapMap project [14]), the beacon would start giving false responses after 6 queries as the budget would be depleted. By using the QI-attack, an attacker would only need 5 queries to determine beacon membership of the victim. Therefore, this countermeasure does not always protect beacon members’ privacy. Accordingly, a query budget that is merely based on the SNPs’ MAFs and that does not consider SNP correlations would fail to protect an individual’s privacy. An attacker using the QI-attack would not exhaust the budget, but still be able to determine the victim’s beacon membership.

Chapter 7

Conclusion

Throughout the course of this thesis, we showed that beacons are sensitive to re-identification attacks. We showed that by including allele frequencies and SNP correlations into the attacker models, the number of queries needed to invade the beacon members' privacy can be significantly reduced. Additionally, we showed that countermeasures that do not consider the MAFs and correlations of SNPs fail to protect the beacon members' privacy. Furthermore, even if individuals apply countermeasures before releasing their genome, such as systematically hiding SNPs with low MAFs, their privacy still could be at stake. As a future work, we therefore need to develop countermeasures that include SNP correlations and allele frequency information to protect sensitive genomic data.

Bibliography

- [1] S. S. Shringarpure and C. D. Bustamante, “Privacy risks from genomic data-sharing beacons,” *The American Journal of Human Genetics*, vol. 97, no. 5, pp. 631–646, 2015.
- [2] J. L. Raisaro, F. Tramr, J. Zhanglong, D. Bu, Y. Zhao, K. Carey, D. Lloyd, H. Sofia, D. Baker, P. Flicek, S. S. Shringarpure, C. D. Bustamante, S. Wang, X. Jiang, L. Ohno-Machado, H. Tang, X. Wang, and J.-P. Hubaux, “Addressing beacon re-identification attacks: Quantification and mitigation of privacy risks,” *The Journal of the American Medical Informatics Association*, vol. 1, no. 1, pp. 1–1, 2016.
- [3] G. M. Church, “The personal genome project,” *Molecular systems biology*, vol. 1, no. 1, 2005.
- [4] F. S. Collins and H. Varmus, “A new initiative on precision medicine,” *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, 2015.
- [5] H. Ledford, “Astrazeneca launches project to sequence 2 million genomes.,” *Nature*, vol. 532, no. 7600, p. 427, 2016.
- [6] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, “Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays,” *PLoS Genet*, vol. 4, no. 8, p. e1000167, 2008.

- [7] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin, “Genomic privacy and limits of individual detection in a pool,” *Nature genetics*, vol. 41, no. 9, pp. 965–967, 2009.
- [8] K. B. Jacobs, M. Yeager, S. Wacholder, D. Craig, P. Kraft, D. J. Hunter, J. Paschal, T. A. Manolio, M. Tucker, R. N. Hoover, *et al.*, “A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies,” *Nature genetics*, vol. 41, no. 11, pp. 1253–1257, 2009.
- [9] P. M. Visscher and W. G. Hill, “The limits of individual identification from sample allele frequencies: theory and statistical analysis,” *PLoS Genet*, vol. 5, no. 10, p. e1000628, 2009.
- [10] D. Clayton, “On inferring presence of an individual in a mixture: a bayesian approach,” *Biostatistics*, p. kxq035, 2010.
- [11] N. Siva, “1000 genomes project,” *Nature biotechnology*, vol. 26, no. 3, pp. 256–256, 2008.
- [12] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, “Addressing the concerns of the lacks family: quantification of kin genomic privacy,” in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 1141–1152, ACM, 2013.
- [13] S. S. Samani, Z. Huang, E. Ayday, M. Elliot, J. Fellay, J.-P. Hubaux, and Z. Kutalik, “Quantifying genomic privacy via inference attack with high-order snv correlations,” in *Security and Privacy Workshops (SPW), 2015 IEEE*, pp. 32–40, IEEE, 2015.
- [14] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch’ang, W. Huang, B. Liu, Y. Shen, *et al.*, “The international hapmap project,” *Nature*, vol. 426, no. 6968, pp. 789–796, 2003.
- [15] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou, “Learning your identity and disease from research papers: information leaks in genome wide association study,” in *Proceedings of the 16th ACM conference on Computer and communications security*, pp. 534–544, ACM, 2009.

- [16] E. Ayday, J. L. Raisaro, J.-P. Hubaux, and J. Rougemont, “Protecting and evaluating genomic privacy in medical tests and personalized medicine,” in *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pp. 95–106, ACM, 2013.
- [17] R. Gorelick and M. D. Laubichler, “Decomposing multilocus linkage disequilibrium,” *Genetics*, vol. 166, no. 3, pp. 1581–1583, 2004.
- [18] Y. Kim, S. Feng, and Z.-B. Zeng, “Measuring and partitioning the high-order linkage disequilibrium by multiple order markov chains,” *Genetic epidemiology*, vol. 32, no. 4, pp. 301–312, 2008.
- [19] S. Feng and S. Wang, “Summarizing and quantifying multilocus linkage disequilibrium patterns with multi-order markov chain models,” *Journal of biopharmaceutical statistics*, vol. 20, no. 2, pp. 441–453, 2010.
- [20] M. M. Al Aziz, R. Ghasemi, M. Waliullah, and N. Mohammed, “Aftermath of bustamante attack on genomic beacon service,” *BMC Medical Genomics*, vol. 10, no. 2, p. 43, 2017.

Appendix A

LRT - Power Calculation

The power $1 - \beta$ of the LRT is determined by the proportion of control individuals for which we can reject the null hypothesis, that is $\Lambda < t_\alpha$. The threshold t_α is found by building the null hypothesis with the 40 case individuals where $\alpha = 0.05$.

For each individual and query x_i we calculate the value of Λ . As t increases the power of the QI-attack shows a zig-zag behavior unlike the Optimal attack and the GI-attack. That is because, as t increases, more queries are needed to determine beacon membership, and more SNPs are inferred in the QI-attack. The more neighbors a posed query can infer from the SNP network, the more extreme the value of Λ changes.

Figure A.1 shows, for three example case and three example test individuals, how Λ steadily decreases for the control individuals and clearly increases for “no” responses of the case individuals (i.e. at queries 24, 26 and 84). For the Optimal attack Λ decreases by a similar value for all individuals that receive a “yes” response, as only one query is asked and the queries have similar MAFs. Therefore, if Control 1 had a lower Λ value at query x_{10} than Case 2, Case 2 will not have a lower Λ value than Control 1 for the following queries, unless Control 1 receives a “no” response (which leads to a significant increase in Λ but is highly unlikely for an individual in the control set). On the contrary, Figure

A.2 shows an irregular behavior of Λ , that is Λ does not steadily decrease as it could be observed for the Optimal attack. This can be explained by the different amount of neighbors in the SNOP network that can be inferred at the different loci. Considering Control 1 and Case 2 again, Control 1 can have a lower Λ value than Case 2 for query position x_{10} . Nevertheless, if query x_{11} of Case 2 has a high amount of neighbors to be inferred from x_{11} and the inferred responses are all “yes” responses, the Λ value of Case 2 decreases significantly and is now lower than the Λ value of Control 1 for x_{11} , as x_{11} of Control 1 had less neighbors in the SNP network.

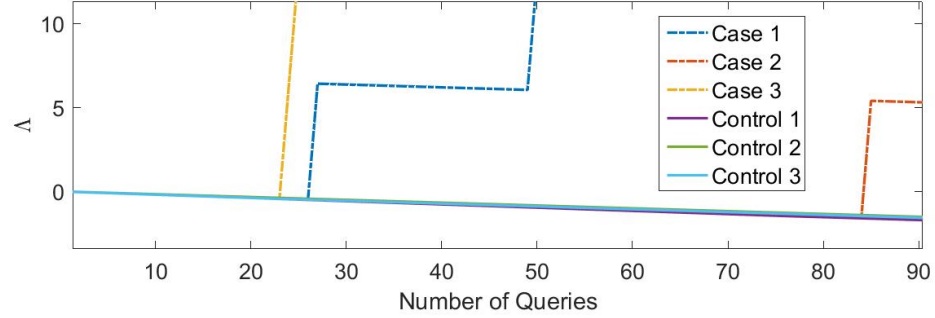


Figure A.1: Example Λ distributions for 3 of the 40 case and 3 of the 20 control individuals of the experiments with a simulated beacon in Section 5.2 for the Optimal attack.

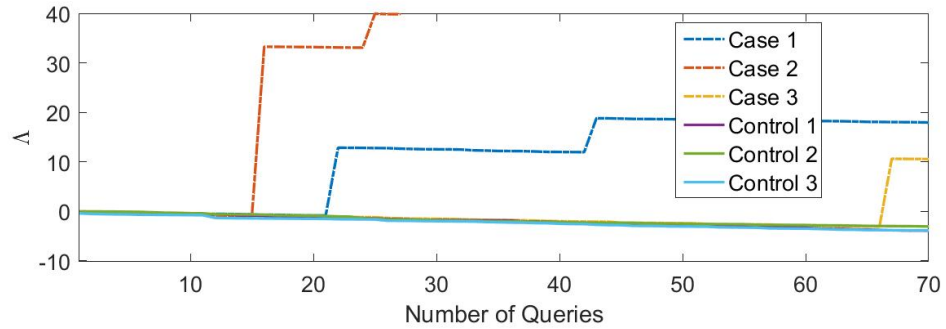


Figure A.2: Example Λ distributions for 3 of the 40 case and 3 of the 20 control individuals of the experiments with a simulated beacon in Section 5.2 for the QI-attack.