# Crimes Cibernéticos no Piauí: Uma Análise Estatística com CRISP-DM e Dados Oficiais (2019–2022)

**Gabriel Linard Leite**[1], **João Paulo de Sá Fonseca**[1], **Matheus Aquino Torquato Reis**[1], **Carlos Eduardo Santos**[1], **Adler Castro Alves**[1], **Maria Hellem Teixeira Abreu**[1]

[1]Curso de Engenharia de Software – Instituto de Ensino Superior – ICEV
Rua Dr. José Auto de Abreu, 2929 – São Cristóvão – Teresina – PI – 64055-260 – Brasil

gabriel.linard@icev.edu.br, joao.paulo@icev.edu.br,
matheus.reis@icev.edu.br,
carlos.santos@icev.edu.br, adler.alves@icev.edu.br,
hellemmaria.mhta@gmail.com

***Abstract.*** *This study analyzes cybercrimes reported in the state of Piauí, Brazil, between 2019 and 2022, based on official data from the Public Security Department and using the CRISP-DM methodology. After data cleaning and categorization with Python libraries (Pandas, Seaborn, and Matplotlib), statistical and visual analyses were applied to identify patterns. The most frequent crimes were digital fraud, device intrusion, and scams, with higher incidence among adults aged 25 to 45. Students, self-employed professionals, and retirees were among the most vulnerable groups. The study highlights digital inequalities and proposes the use of predictive models and data-driven public policies to strengthen information security.*

***Resumo.*** *Este estudo analisa crimes cibernéticos registrados no Piauí entre 2019 e 2022, com base em dados oficiais da Secretaria de Segurança Pública e na metodologia CRISP-DM. Após o tratamento e categorização com bibliotecas Python (Pandas, Seaborn e Matplotlib), aplicaram-se análises estatísticas e visuais para identificar padrões. Os crimes mais recorrentes foram estelionato digital, invasão de dispositivos e fraudes, com maior incidência entre adultos de 25 a 45 anos. Estudantes, autônomos e aposentados figuram entre os grupos mais vulneráveis. O estudo destaca desigualdades digitais e propõe o uso de modelos preditivos e políticas públicas orientadas por dados para fortalecer a segurança da informação.*

## 1. Introdução

Nas últimas décadas, o avanço tecnológico transformou as relações sociais, econômicas e institucionais, criando oportunidades — mas também novos vetores de risco. Destacam-se os crimes cibernéticos, que envolvem fraudes eletrônicas, invasões de dispositivos e ataques a infraestruturas críticas, desafiando os modelos tradicionais de segurança pública e exigindo abordagens inovadoras baseadas em dados [Provost et al. 2013, Han et al. 2012].

No Brasil, a promulgação da *Lei n° 12.737/2012*, conhecida como *Lei Carolina*

*Dieckmann*, representou um marco no enfrentamento aos crimes digitais, ao tipificar condutas como invasão de dispositivos, alteração de dados e divulgação não autorizada de informações pessoais [Baesens et al. 2014]. Apesar disso, a crescente sofisticação tecnológica tem ampliado a complexidade das infrações e exposto a fragilidade dos mecanismos tradicionais de prevenção — especialmente em estados com infraestrutura digital emergente, como o Piauí [Anderson et al. 2009].

Diante desse cenário, este estudo realiza uma análise sistemática dos crimes cibernéticos registrados no estado do Piauí entre 2019 e 2022, com base em dados oficiais de órgãos de segurança pública. A base reúne variáveis como sexo, idade, profissão, cidade, bairro e natureza da ocorrência, permitindo um processo estruturado de limpeza, padronização e transformação dos dados, com tratamento de valores ausentes, categorização de faixas etárias e correção de inconsistências textuais [McKinney et al. 2018].

A investigação adota a metodologia *CRISP-DM* (*Cross Industry Standard Process for Data Mining*), reconhecida por sua aplicação em ciência de dados e mineração de conhecimento. Essa abordagem estruturada guiou todo o fluxo analítico — da compreensão do problema à comunicação dos resultados [Fayyad et al. 1996] — e mostrou-se eficaz diante da fragmentação dos dados sobre cibercriminalidade.

Com base nesse modelo, este trabalho utiliza a linguagem *Python* e bibliotecas como *Pandas*, *Seaborn* e *Matplotlib* para operações de agregação, filtragem, visualização estatística e construção de gráficos descritivos e multivariados [McKinney et al.2018, Hunter et al.2007, Waskom et al.2021]. A próxima seção detalha essa estrutura metodológica e o processo analítico adotado para fundamentar os resultados e recomendações do estudo.

## 2. Metodologia

A análise dos crimes cibernéticos neste estudo foi conduzida com base no modelo *CRISP-DM* (*Cross Industry Standard Process for Data Mining*), que estrutura logicamente as etapas de preparação, exploração e interpretação dos dados [Fayyad et al. 1996]. A base é composta por registros reais de crimes, contendo variáveis como tipo de ocorrência, idade, sexo, raça/cor, profissão e localidade das vítimas. A inspeção inicial foi realizada com funções da biblioteca *Pandas* [McKinney et al. 2018], permitindo identificar a estrutura e qualidade dos dados.

O pré-processamento envolveu limpeza, padronização textual, tratamento de valores nulos e categorização de faixas etárias para facilitar comparações. A análise exploratória de dados (*EDA*) utilizou gráficos de barras, histogramas, mapas de calor e tabelas de contingência com o apoio das bibliotecas *Seaborn* e *Matplotlib* [Hunter et al. 2007, Waskom et al. 2021].

As visualizações revelaram padrões relevantes, como maior incidência de crimes como estelionato e invasão de dispositivos (*art.154-A*) entre adultos de 25 a 45 anos, especialmente entre vítimas autônomas, estudantes e aposentados [Kelleher et al. 2015, Rubinstein-Kroese 2016]. Os cruzamentos entre sexo e raça/cor indicaram maior vitimização de pessoas pardas, sugerindo desigualdades estruturais no ambiente digital.

| Coluna | Descrição |
|---|---|
| natureza_ocorrencia | Tipo de crime cibernético registrado (ex: estelionato, invasão de sistema) |
| sexo | Sexo da vítima (masculino, feminino, não informado) |
| idade | Idade da vítima em anos |
| faixa_etaria | Faíxa etária categorizada a partir da idade (ex: 18-25, 3 |
| raca_cor | Raça ou cor da pele da vítima (ex: preta, branca, parda) |
| profissão | Ocupação declarada da vítima |
| cidade_vítima | Cidade onde ocorreu a ocorrência |
| bairro_vitima | Bairro de residência da vítima |
| baírro_ocorrencia | Bairro onde ocorreu o crime (caso diferente do residên- |
| data_ocorrencia | Data do registro do crime |

**Figure 1. Dicionário de variáveis utilizadas na análise dos crimes cibernéticos. Fonte: SSP, dados tratados pelos autores.**

Apesar do caráter descritivo e exploratório, os resultados indicam o potencial de uso de modelos preditivos futuros, como classificadores de risco por perfil. A abordagem metodológica — unindo estatística descritiva, programação em *Python* e estruturação lógica — permitiu não só descrever os dados, mas também gerar *insights* relevantes ao contexto brasileiro, especialmente em regiões com inclusão digital desigual. Tais achados podem orientar políticas públicas voltadas à educação e prevenção digital, fortalecendo a cibersegurança baseada em evidências.

## 3. Desenvolvimento de Pesquisa

A pesquisa baseou-se em uma base de dados estruturada com variáveis relacionadas à natureza do crime, sexo, idade, raça/cor, profissão e localidade da vítima. Essas informações foram tratadas para identificar padrões de vitimização e relações entre atributos sociodemográficos e a recorrência de delitos cibernéticos.

A etapa inicial incluiu inspeção e pré-processamento com as bibliotecas *Pandas* e *Seaborn* [McKinney et al. 2018, Waskom et al. 2021], com verificação da qualidade dos dados, remoção de duplicatas, tratamento de ausências e normalização textual. Variáveis com excesso de dados nulos foram descartadas, enquanto atributos categóricos — como "sexo", "raça_cor" e "profissão" — passaram por padronização. A variável "idade" foi convertida para tipo numérico e segmentada por faixas etárias.

A análise estatística descritiva considerou medidas de tendência central e dis-

persão (média, mediana, desvio-padrão e quartis). Foram utilizadas visualizações como gráficos de barras, histogramas e mapas de calor, que permitiram observar associações relevantes. Representações mostraram que vítimas jovens adultas foram mais comuns em estelionato e invasão de dispositivos, enquanto idosos apresentaram maior exposição a fraudes financeiras — um padrão vinculado às desigualdades de acesso e letramento digital.

Algumas análises foram aprofundadas com a amostragem de Monte Carlo [Rubinstein-Kroese 2016], simulando distribuições empíricas e testando a robustez de proporções entre categorias (ex.: tipos de crime por sexo ou profissão). Apesar do foco descritivo, a base também permitiu testes de associação (qui-quadrado) e modelos de regressão, úteis para futuros estudos preditivos. A sistematização do processo analítico garantiu replicabilidade metodológica em diferentes contextos, permitindo adaptações conforme as características locais. Com isso, foi possível gerar interpretações significativas e orientar políticas públicas voltadas à segurança da informação, tanto em nível regional quanto nacional [Provost et al. 2013, Bishop et al. 2006].

## 4. Análise

A análise dos crimes cibernéticos foi realizada a partir de transformações e visualizações aplicadas ao dataset, com uso das bibliotecas *Matplotlib, Seaborn e Pandas* [McKinney et al. 2018, Waskom et al.2021]. Foram elaborados gráficos descritivos e mapas de calor para identificar padrões de vitimização por idade, sexo, raça, profissão e natureza do crime.

Inicialmente, na *Figura 2*, os dados revelaram que a faixa etária entre 25 e 45 anos concentra o maior número de vítimas, com destaque para os 30 a 35 anos, faixa economicamente ativa e exposta a riscos digitais como transações e redes sociais [Hunter et al. 2007, Fayyad et al.1996].

As análises também apontaram um leve predomínio de vítimas do sexo feminino em determinados tipos de crimes, especialmente fraudes afetivas e estelionato digital. Para evitar generalizações, essa conclusão baseou-se na contagem total por categoria (`value_counts`) e visualizações cruzadas com `pd.crosstab()` e `sns.heatmap()`, como mostrado na *Figura 3*. Testes qui-quadrado de independência poderão ser utilizados em estudos futuros para validar estatisticamente essas associações.

Ao cruzar os dados de sexo e raça/cor, foi possível observar que pessoas pardas foram as principais vítimas, independentemente do gênero, seguidas por brancos e pretos. Esse padrão sugere que desigualdades raciais estruturais também se refletem no ambiente digital, como mostra a *Figura 4*, alinhando-se à literatura sobre desigualdades tecnológicas [Bishop et al.2006, Anderson et al.2009].

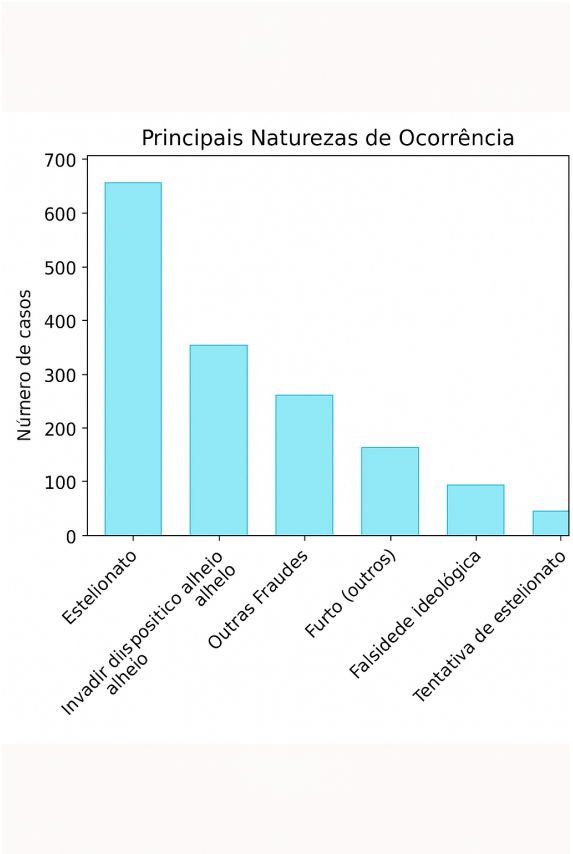


**Figure 2. Gráfico de Crimes mais Frequentes.** ***Fonte: SSP, dados tratados pelos autores.***

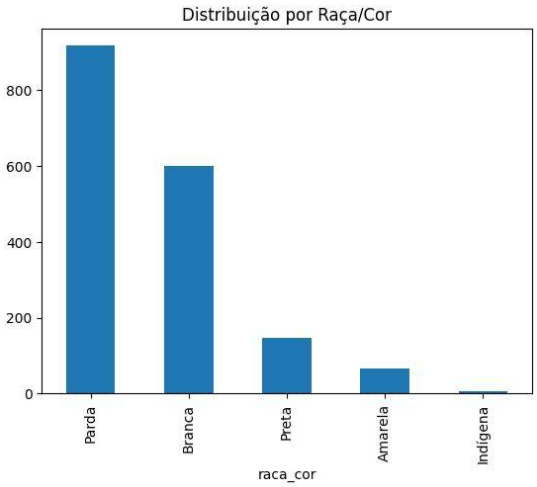


**Figure 3. Gráfico de Barras: Raça/Cor Vítimas.** ***Fonte: SSP, dados tratados pelos autores.***

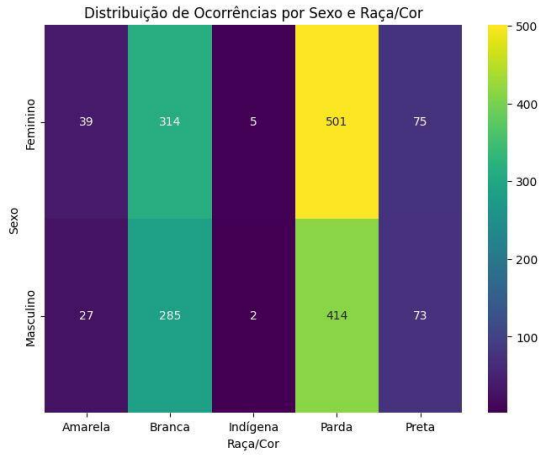


**Figure 4. Relação das Vítimas pelo Sexo.** ***Fonte: SSP, dados tratados pelos autores.***

Quanto à *Figura 5*, que representa a categoria voltada à profissão, estudantes, autônomos, aposentados e servidores públicos foram os mais afetados. Estudantes (18–25

anos) mostraram maior exposição a crimes como phishing e engenharia social, enquanto aposentados e servidores foram vítimas frequentes de fraudes financeiras [Kelleher et al.2015, Baesens et al.2014].

Além da visualização direta das variáveis, foi aplicado um *heatmap* condicional (Figura 6), evidenciando que fraudes de identidade e invasão de dispositivos ocorrem mais em adultos de meia-idade. Para medir impacto, propôs-se uma taxa de vitimização relativa (não aplicada neste estudo por falta de dados populacionais):

$$\text{Taxa de Vitimação} = \left( \frac{\text{Número de ocorrências em um grupo}}{\text{População total do grupo}} \right) \times 1000 \qquad (1)$$

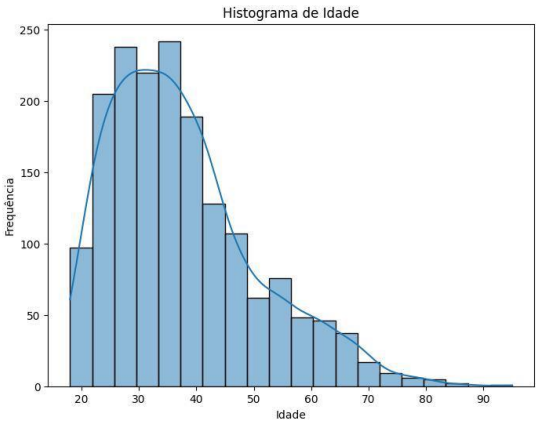


**Figure 5. Gráfico de Barras das Vítimas por Idade. *Fonte: SSP, dados tratados pelos autores.***


**Figure 6. Gráfico dos Principais Crimes Cibernéticos. *Fonte: SSP, dados tratados pelos autores.***

Reconhecem-se limitações como subnotificação, viés de registro e sub-representação de populações com menos acesso à internet, o que afeta a confiabilidade dos dados. Como continuidade, recomenda-se o uso de abordagens preditivas com regressão logística, curvas *ROC* e *MLE (Maximum Likelihood Estimation)* para estimar perfis de risco [Dean-Ghemawat 2008, Rubinstein-Kroese 2016].

A distribuição geral das ocorrências por sexo evidenciou maior vitimização feminina em fraudes afetivas e invasões de privacidade [Kelleher et al.2015, Hastie et al.2009 ]. A *Figura 7* apresenta essa distribuição, confirmando um volume levemente superior de vítimas do sexo feminino — dado que reforça achados prévios.

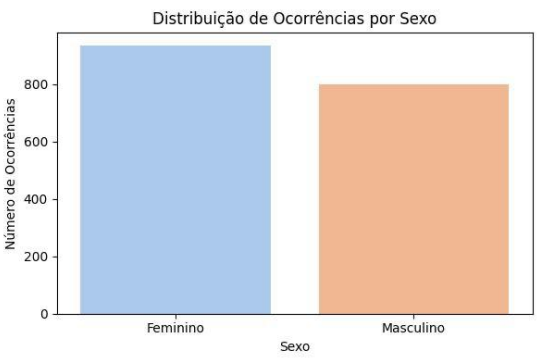


**Figure 7. Relação entre Gênero e Etnia na Exposição de Crimes Digitais. *Fonte: SSP, dados tratados pelos autores.***

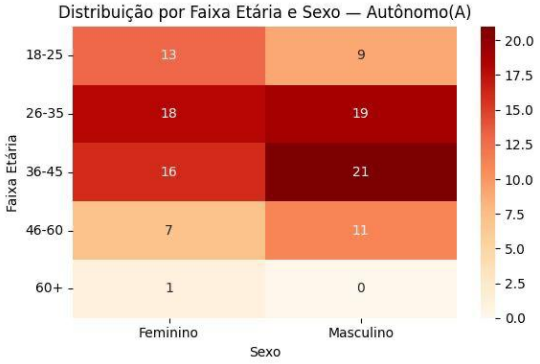


**Figure 8. Relação entre Faixa Etária, Sexo e Tipo de Crime. *Fonte: SSP, dados tratados pelos autores.***

A análise da variável profissão trouxe novos elementos. As profissões com maior incidência foram estudantes, autônomos, aposentados, professores e médicos [Kelleher et al.2015]. Isso sugere que os crimes cibernéticos afetam múltiplos setores, conforme o tipo de atividade realizada online [Baesens et al.2014].

A *Figura 9* comprova que estudantes concentram-se na faixa de 18–25 anos, enquanto aposentados predominam acima dos 60. Autônomos, professores e técnicos se destacam entre 26 e 45 anos [Kelleher et al.2015, Baesens et al.2014].

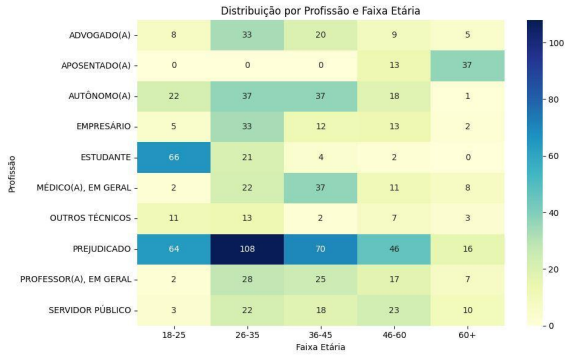


**Figure 9. Relação entre a Profissão x Faixa Etária das Vítimas. *Fonte: SSP, dados tratados pelos autores.***

Essas últimas visualizações enriquecem a compreensão do fenômeno ao incorporar múltiplas variáveis de forma simultânea, promovendo uma análise mais robusta e aplicável à formulação de estratégias de cibersegurança orientadas por evidência, fortalecidas por abordagens preventivas baseadas em dados: Google Colab.

## 5. Conclusão

A aplicação da metodologia *CRISP-DM* possibilitou um fluxo analítico estruturado para investigar crimes cibernéticos ocorridos entre 2019 e 2022. Com base em dados

reais, foi possível extrair informações relevantes sobre vítimas, tipos de crime e perfis mais vulneráveis.

O uso integrado de bibliotecas como *Pandas*, *Seaborn* e *Matplotlib* permitiu desde a limpeza dos dados até a construção de visualizações multivariadas, como histogramas, mapas de calor e gráficos de barras. Identificaram-se padrões entre variáveis como faixa etária, profissão, raça/cor e tipo de delito, com destaque para vítimas entre 26 e 45 anos envolvidas em crimes como estelionato e invasão de dispositivos eletrônicos, especialmente entre autônomos, estudantes e aposentados [figura 9].

A análise também evidenciou limitações, como a ausência de variáveis sobre escolaridade, renda e tempo de exposição online, dificultando a compreensão das desigualdades de raça e gênero. A subnotificação e o viés de registro, comuns nesses crimes, comprometem a representatividade de casos como calúnia e divulgação de conteúdo íntimo, que muitas vezes não são formalmente denunciados.

Apesar do foco geográfico no Piauí, os achados podem ser replicados em outras regiões com estrutura de dados semelhante, viabilizando estudos comparativos sobre crimes digitais em diferentes contextos. A promulgação da *Lei n° 12.737/2012* — conhecida como *Lei Carolina Dieckmann* — foi um marco legal, mas os dados mostram que a ocorrência desses crimes permanece alta, exigindo integração entre investigação, educação digital e fiscalização [Han et al.2012].

Para avanços, propõe-se o uso de modelos estatísticos mais robustos, como regressão logística e testes de hipótese, além de técnicas como *MLE* (*Maximum Likelihood Estimation*) e curvas *ROC*, que permitem estimar riscos e validar classificadores.

Em síntese, o estudo reforça o papel da ciência de dados na segurança digital, destacando a importância da estruturação analítica e da visualização como ferramentas para orientar políticas públicas e proteger comunidades vulneráveis.

## Referências


[Provost et al. 2013] Provost, Foster; Fawcett, Tom. (2013). *Data science for business*. Sebastopol: O'Reilly Media.

[Han et al.2012] Han, Jiawei; Kamber, Micheline; Pei, Jian. (2012). *Data mining: concepts and techniques*. 3rd ed. San Francisco: Morgan Kaufmann.

[Witten et al.2011] Witten, Ian H.; Frank, Eibe; Hall, Mark A. (2011). *Data mining: practical machine learning tools and techniques*. 3rd ed. San Francisco: Morgan Kaufmann.

[Kelleher et al.2015] Kelleher, John D.; Mac Namee, Brian; D'Arcy, Aoife. (2015). *Fundamentals of machine learning for predictive data analytics*. Cambridge: MIT Press.

[Hastie et al.2009] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. (2009). *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: Springer.

[Bishop et al.2006] Bishop, Christopher M. (2006). *Pattern recognition and ma-*

*chine learning*. New York: Springer.

[Dean & Ghemawat 2008] Dean, Jeffrey; Ghemawat, Sanjay. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

[Murphy et al. 2012] Murphy, Kevin P. (2012). *Machine learning: a probabilistic perspective*. Cambridge: MIT Press.

[McKinney et al. 2018] McKinney, Wes. (2018). *Python for data analysis*. 2nd ed. Sebastopol: O'Reilly Media.

[Hunter et al. 2007] Hunter, John D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

[Waskom et al.2021] Waskom, Michael L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60).

[Box & Draper 1987] Box, George E. P.; Draper, Norman R. (1987). *Empirical model-building and response surfaces*. New York: Wiley.

[Rubinstein & Kroese 2016] Rubinstein, Reuven Y.; Kroese, Dirk P. (2016). *Simulation and the Monte Carlo method*. 3rd ed. New York: Wiley.

[Fayyad et al.1996] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54.

[Baesens et al.2014] Baesens, Bart. (2014). *Analytics in a big data world: the essential guide to data science and its applications*. Hoboken: Wiley.

[Anderson et al.2009] Anderson, Ross; Moore, Tyler; Acquisti, Alessandro, et al. (2009). Information security economics—and beyond. In: Acquisti, A. et al. (Eds.). *Economics of information security and privacy*. New York: Springer, p. 17–44.