Universidade Federal de Pernambuco – Campus Recife

Centro de Informática – 2018.1

Aluno: Marcos da Silva Barreto

Unique in the Crowd: The privacy bounds of human mobility

Os autores analisaram dados de mobilidade humana e constataram que o "rastro" deixado pelos usuários de operadoras de telefonia é altamente característico e capaz de identificá-los de forma única. A base de dados analisada era anônima de uma empresa de telefonia também não identificada. Os registros eram datados entre o período de abril de 2006 e junho de 2007 para uma média de 1.5 milhões de usuários da parte ocidental de um país europeu. Nessa base, os dados consistiam do momento em que o usuário interagiu com a operadora (ligação ou SMS) e o local da antena a qual ele se conectou. Os dados então são pontos espaço-temporais.

O objetivo dos autores é avaliar quantitativamente a probabilidade de identificação mesmo em base de dados simplesmente anônima, segundo os padrões de comportamento ou "rastros". Para isso foi tomado um conjunto de p pontos espaço-temporais aleatórios (I_p), e avaliar o valor de ε, a singularidade do "rastro", ou seja, se a partir da base de dados completa existe apenas um subconjunto $S(I_p)$ que equivale exatamente aos p pontos de I_p, $|S(I_p)| = 1$. Por exemplo, é conhecido que uma pessoa realizou uma ligação às 6 da manhã na rua A e outra às 8 da noite na rua B, buscando esses p = 2 pontos na base e encontrando apenas um subconjunto $S(I_p)$, ela é identificável na base, e assim é possível conhecer todo o seu histórico.

Para representar espacialmente os dados os autores utilizaram um "mapa" com a distribuição das antenas ao longo do território analisado, onde cada área é a metade da distância máxima entre duas antenas. Esse mapa é semelhante a um diagrama (malha) de Voronoi, onde cada região é a área de cada antena. Com essa configuração a resolução dos pontos se dá por uma antena e um intervalo de uma hora. Após a análise da estratégia comentada anteriormente foi encontrado que 4 pontos escolhidos aleatoriamente era o suficiente para caracterizar unicamente 95% ($\varepsilon > 0.95$) dos usuários, e com apenas 2 pontos alcança a marca de 50% ($\varepsilon > 0.5$).

Porém o valor de ε depende da resolução espaço-temporal da base de dados. Para avaliar isso os autores reduziram a resolução fazendo agregações. Clusterizaram antenas, aumentando as regiões do mapa de Voronoi e aumentaram a janela de tempo de observação. Assim, para 4 pontos em resolução de 5 horas e 5 antenas, o valor de ε foi de apenas 50%. Porém essa restrição pode ser contornada por aumentar o número de p. Mesmo diminuindo a resolução dos dados, tornando a granularidade alta, não torna a base de dados segura.

Unique in the shopping mall: On the reidentifiability of credit card

Esse artigo segue as premissas do artigo anterior. A base de dados em questão são os metadados dos cartões de crédito. Os autores afirmam que as onipresenças das tecnologias atualmente criam metadados pessoais em uma escala gigantesca. Esses dados têm grande potencial e muitos benefícios, em diversas áreas desde saúde a publicidade, porem estão condicionados a sua disponibilidade, e torna-los visíveis requer uma enorme garantia do risco de reidentificação e o entendimento dos limites da privacidade. Essas bases assemelham às discutidas anteriormente sobre não conter informações de identificação, como nomes, números e endereços.

O objetivo dos autores é avaliar a probabilidade de identificação nesses metadados de cartões de crédito. A base de dados anônima contém metadados de aproximadamente 1.1 milhões de pessoas em um país da OECD (*Organisation for Economic Co-operation and Development*) coletados durante o período integral entre janeiro e março. Dentre os dados analisados, haviam informações acerca do gênero e nível de renda. Observando que mulheres e pessoas de alta renda são mais fáceis de reidentificar do que os demais.

Nos metadados, cada transação armazena pelo menos o dia, a loja, e o preço pago. Para o objetivo do artigo, os autores utilizaram a mesma noção de singularidade (ε) do artigo anterior, não necessitando a explicação do método. Os pontos espaço-temporais inicialmente considerados foram o dia e a loja. Após análise, 90% dos usuários são reidentificados com apenas 4 pontos ($\varepsilon > 0.9$). Posteriormente notaram que o preço da transação poderia ser utilizado para a reidentificação, agora o ponto passa a ser uma tripla (dia, loja, preço). Para o estudo, o valor exato do preço passou a ser categorizado por intervalos de preços, com resolução conhecida por (alfa), inicialmente (alfa) = 0.5. Com isso concluíram que com o preço houve um aumento, em média, de 22% no caráter de reidentificação.

A resolução dos dados também foi avaliada, as lojas foram agrupadas de acordo com sua distância, a janela de tempo aumentada de 1 para 15 dias e o tamanho dos intervalos de preços alterando (alfa) para 0.75. Porém aumentar a granularidade não é suficiente para proteger a privacidade, sendo facilmente superado adicionando pontos, como observado no outro trabalho.