



これらの駅を経由している乗客は あなた 1 人しかいない

菊池 浩明 明治大学総合数理学部

〔受賞論文〕

Zipf Distribution Model for Quantifying Risk of Re-identification from Trajectory Data

Hiroaki Kikuchi (School of Interdisciplinary Mathematical Sciences, Meiji University), Katsumi Takahashi (NTT Secure Platform Laboratories)

Journal of Information Processing Vol.24, No.5, pp.816-823 (2016)

関東地方の4千万人の乗客が2,500種類の駅から駅への乗降を繰り返している。これらの乗降の履歴をいくつ集めると1人を識別することができるだろうか？

最も乗降が多いのは、新宿駅である。2015年1日平均76万の乗降があり、これは乗降履歴全体の9%に及ぶ。電車に乗ったら10回に1回は新宿駅を使っていることになり、新宿駅の履歴があってもごく当たり前すぎて個人を識別するには役に立たない。第100位は阿佐ヶ谷駅で、新宿駅の6%しか乗降がないが、それでも延べ4万4千人がこの駅を使う。あきた白神駅は27人だそうだ。これならば、この駅の履歴から候補が絞れる。

という問題をくそ真面目に研究した論文である。私たちは「鉄ちゃん」ではないし、時刻表アリバイを崩すミステリーマニアというわけでもない。この研究の背景には、2013年当時、ビッグデータの導入要請を受けて改定が検討されていた個人情報保護法がかかっている。私(菊池)と高橋は、内閣官房IT総合戦略室 パーソナルデータに関する検討会技術検討ワーキンググループ・構成員として、霞が関ビルの裏側にある尚友会館に通い続ける日々であった。従来法の個人情報の範囲を拡大するのか、グレーにあたる部分に新たな概念を定義するか、このWGに与えられた大きな課題の中でも、とりわけ、交通系ICカードに蓄積されている乗降履歴を安全に匿名化することは難問であった。乗降履歴にどれだけの再識別リスクがあるのか、定量化しようにもデータの

入手さえできなかった。

そこで、いくつもの仮定を与えて乗降の振舞いを単純なZipf則による数理モデルで近似したものが本論文の下地になっている。本論文の主要な結論は、モデルで算出した再識別不能確率(定理3.2)と匿名化に必要な削除レコード数の下限(定理3.3)である。このモデルでは、高々3駅分の履歴から98%の乗客が一意に識別される。乗降履歴の再識別リスクは、我々の想像以上に高いものであった。

そんなにリスクが高いはずがない。何か間違っているに違いない。

そう不思議に思ってもらえたら嬉しい。この拙作が、モデルの改善やより正確な評価方法の提案につながり、個人情報の保護と活用のための議論が進むきっかけになることを願っている。

(2017年5月17日受付)

菊池 浩明 (正会員) kikn@meiji.ac.jp

1990年明治大学院博士前期課程修了。1994年同博士(工学)。(株)富士通研究所、東海大学情報通信学部を経て、2013年より明治大学総合数理学部先端メディアサイエンス学科教授。1990年日本ファジィ学会奨励賞、1993年本会奨励賞、1996年SCIS論文賞、2010年本会JIP Outstanding Paper Award。2013年IEEE AINA Best Paper Award。電子情報通信学会、日本知能情報ファジィ学会、IEEE、ACM各会員。本会フェロー。

高橋 克巳 (正会員) takahashi.katsumi@lab.ntt.co.jp

1988年東京工業大学卒業。2006年東京大学大学院博士課程修了。1988年日本電信電話(株)入社。NTTセキュアプラットフォーム研究所主席研究員。2011年筑波大学客員教授。2012年総務省統計研究所客員研究官。2016年国立情報学研究所客員教授。本会では2010年情報規格調査会規格役員。2013年理事。論文賞(2000年、2011年)。博士(情報理工学)。本会フェロー。