# Modelling data flows as graphs to apply user privacy constraint

Abhishek Tiwari

Citation: *A. Tiwari*, "Modelling data flows as graphs to apply user privacy constraint", Abhishek Tiwari, 2024. doi:10.59350/j1yfp-wrq24

Published on: September 08, 2024

Personal data processing forms the backbone of many big tech service providers. Tech giants like Netflix, Meta, and Amazon employ intricate networks of microservices to automatically process user data, creating complex data flows that span multiple layers of computation. Privacy concerns and regulatory frameworks like the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), Brazilian Lei Geral de Protec¸a˜o de Dados (LGPD), Digital Markets Act (DMA) mandate that users should have control over usage and processing of their data. Users can opt out of certain types of data processing. These user-imposed constraints can significantly impact the service provider's utility.

The crux of the problem lies in the complexity of modern data processing systems. When you post a photo on social media or make a purchase online, your data doesn't simply go from point A to point B. Instead, it flows through a labyrinth of microservices, each performing specific functions and passing data along. This intricate web of data flows has made implementing user privacy preferences a complex distributed computing problem. The challenge, therefore, lies in respecting user privacy preferences while minimizing the loss of utility for service providers. This delicate balance must be achieved within the context of increasingly complex and interconnected data processing systems.

The journey of personal data typically begins at the front-end applications, which dispatch requests to various services. Each service performs a specific business function, often generating predictions or inferences that feed into other services. For example, in a social media platform, a user's location data might first be used to process profile information, then sent to usage analytics and group suggestion services. The insights from these services could then inform ad ranking and product recommendation services. This intricate web of data processing serves crucial business goals. Ad ranking services drive revenue through personalized advertisements, while product recommendation services may generate commissions on sales. The ultimate aim is to maximize the service provider's utility. When a Users opt out of certain types of data processing, such as preventing their location data from being used for personalized ads or product recommendations, users privacy preferences must be respected. For tech companies, these user-imposed privacy constraints poses significant challenges .

The scale of modern computing systems further complicates this issue because data processing is typically not performed in isolation but through pipelines, workflows, or a network of microservices. In reality, data flows may involve hundreds of nodes across numerous processing stages, steps or microservices. Meta's microservice topology, for instance, encompasses over 12 million service instances with more than 180,000 communication edges. In such vast systems, restricting data flow to certain services can have far-reaching effects, potentially degrading the quality of inferences used by other services and, consequently, the overall utility for the service provider.

## Graph Theory for Consent Management

A research paper by Filipczuk et. al proposes a new approach to automatically satisfy fine-grained privacy constraints of a user in a way which also optimises the service provider's gains from processing. Study models these complex data flows as graphs. In this approach, the stages of data processing become vertices, and the flow of data between them becomes edges. User privacy constraints are represented as pairs of vertices that need to be disconnected. This elegant abstraction transforms a messy, real-world problem into a well-defined mathematical one, opening up a whole new toolbox for tackling privacy management.

By framing the problem as finding a subgraph that satisfies all user constraints while maximising utility for the service provider, the researchers have created a framework that acknowledges the legitimate needs of both users and businesses. This is crucial because, let's face it, businesses aren't going to adopt privacy measures that cripple their ability to function.

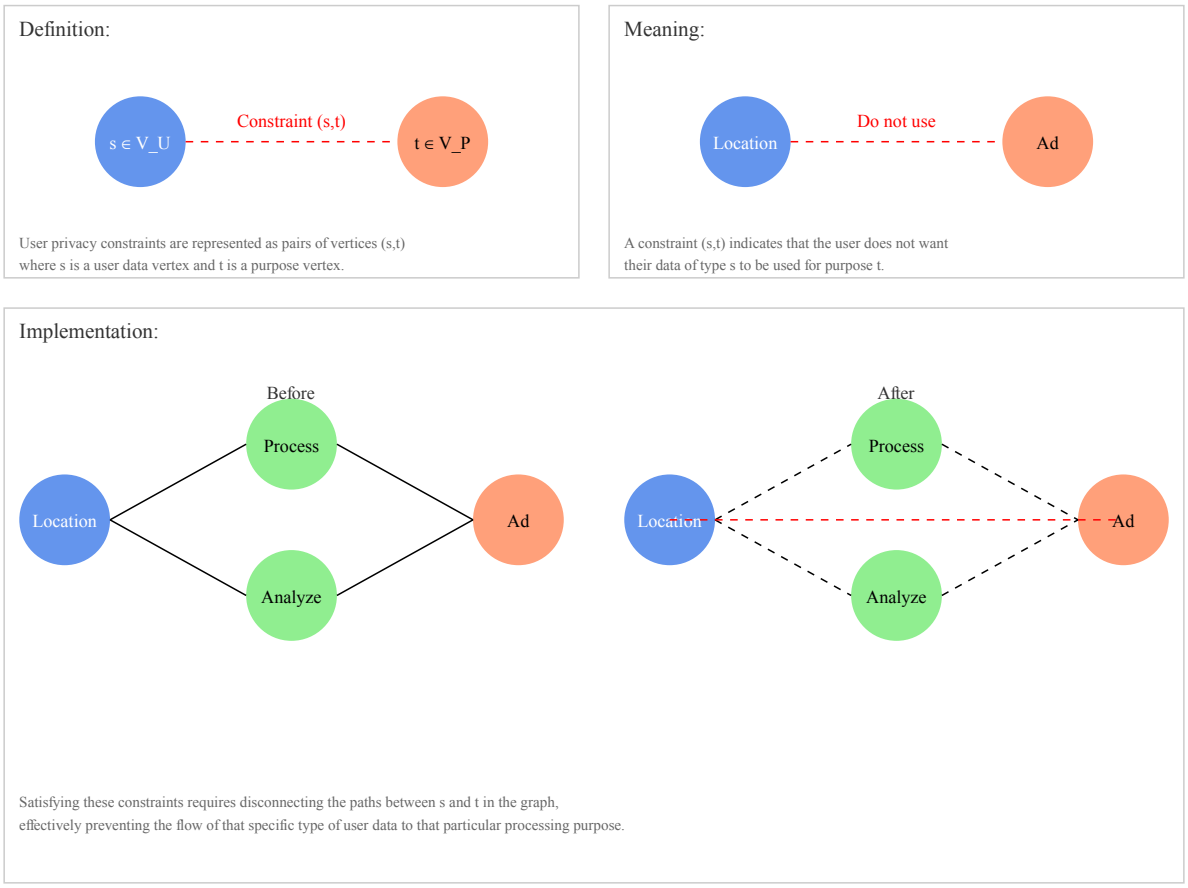


**Figure 1:** Consented Data Workflow problem (CDW): given user constraints expressed in terms of the vertices that they do not wish to be connected, find a subgraph of the original workflow where these constraints are satisfied.

The study introduces several algorithms for solving this problem, each with its own strengths and weaknesses. The standout performer is RemoveMinMC, based on the minimum multicut problem. This algorithm can handle graphs with thousands of nodes and tens of constraints in just seconds, all while providing near-optimal solutions. For a problem as complex as privacy management in large-scale systems, this level of performance is nothing short of remarkable.

Moreover, this graph-based model allows for highly specific privacy constraints, giving users granular control over their data. Want to allow your location data to be used for restaurant recommendations but not for targeted ads? This model can handle that level of specificity.

## Challenges

Of course, no solution is without its challenges. The researchers highlight several areas that need further work. For instance, accurately modeling real-world data flows as graphs is still a significant hurdle. We'll need automated tools to generate these graphs from existing systems if this approach is to be widely adopted. Additionally, the current study uses a simplified model for data utility. In the real world, the value of data is often more complex and context-dependent. Developing more sophisticated utility models will be crucial for the practical application of this approach.

Perhaps the most intriguing challenge is scaling this solution to handle millions of users, each with their own privacy preferences. The researchers suggest exploring methods for clustering similar user types or incrementally updating solutions. This touches on a fundamental question in privacy management: how do we balance personalised control with system-wide efficiency?

## Conclusion

I believe this research represents a significant step forward in our ongoing struggle to balance privacy and utility. Paper also opens the way for many follow-up research contributions. By providing a mathematical framework for this balancing act, it offers a way to harmonize privacy protection with the realities of modern data processing, will be essential.