Flow Provenance in Temporal Interaction Networks

Chrysanthi Kosyfaki

Department of Computer Science & Engineering, University of Ioannina, Greece

xkosifaki@cs.uoi.gr

Xi'an, Shaanxi, China SIGMOD 2021 /PODS

Introduction

- Many real-world applications can be represented as temporal interaction networks (TINs), which capture the information flow between entities over time.
- Previous work mainly focused on developing systems for efficiently provenance storage or analyzing workflow graphs.
- Does not consider the quantities which are transferred among the vertices.

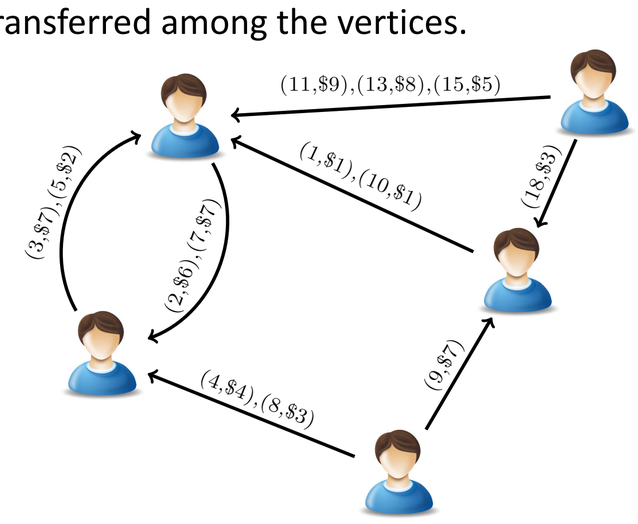


Figure 1: An example of a Bitcoin Network

- **Goal:** study and define the provenance in TINs and track the origin of the quantities that are accumulated at the vertices over time.
- **Applications:** social networks, communication networks, road networks, financial networks etc.

Background

- Data provenance is a core concept in database query evaluation and workflow graphs.
- In query evaluation, for example, it is important to know which data in the database contribute to a query result.
- Provenance can also be defined from two different perspectives: where and why provenance:
  - **why-provenance** finds the entities in the query evaluation plan that contribute to the result.
  - **where-provenance** finds the tuples in the source tables of the query that contribute to the result.

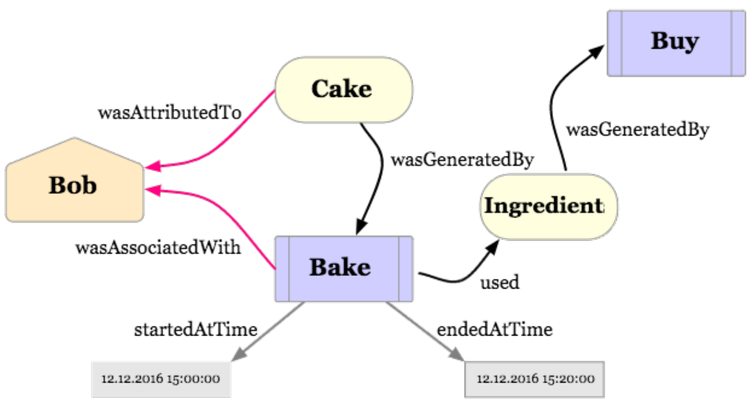


Figure 2: An example of workflow provenance

Our contributions are the following:

- study and define a flow provenance problem in temporal interaction networks
- propose different models based on realistic assumptions
- conduct experiments using real data

Problem Definition

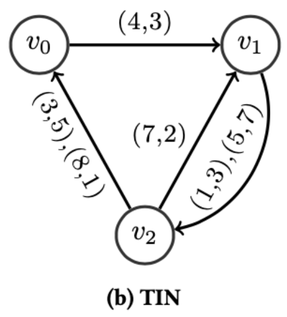
- **Input Graph**

The input graph to our problem is a directed graph G(V,E,R) where each edge (v,u) in E captures the history of interactions from vertex v to vertex u. R denotes the set of interactions on all edges of E.

Each interaction r ∈ R is characterized by a quadruple <r.s, r.d, r.t, r.q> where r.s is the source (destination) vertex of the interaction, r.t is the time when the interaction took place and r.q is the transferred quantity from vertex r.s to r.d due to interaction r.

r.s	r.d	r.t	r.q
v_1	v_2	1	3
v_2	v_0	3	5
v_0	v_1	4	3
v_1	v_2	5	7
v_2	v_1	7	2
v_2	v_0	8	1

(a) Interactions



(b) TIN

Figure 3: A set of interactions and a TIN

- **Provenance Problem:** Given a TIN G(V,E,R) at any time t and at any vertex v ∈ V, find the origins of the quantity, which is accumulated at v.

Models

Three different relay models, which are based on realistic assumptions:

- **Least recently born selection model:** this model gives priority to the quantities that have the oldest birth timestamps. Any generated quantity should be marked with the vertex v that generates it and the timestamp t when it is generated.

r.s	r.d	r.t	r.q	B_{v_0}	B_{v_1}	B_{v_2}
v_1	v_2	1	3	∅	∅	{(1,1,3)}
v_2	v_0	3	5	{(1,1,3),(2,3,2)}	∅	∅
v_0	v_1	4	3	{(1,1,2)}	{(1,1,1),(2,3,2)}	∅
v_1	v_2	5	7	{(1,1,2)}	∅	{(1,1,1),(2,3,2),(1,5,4)}
v_2	v_1	7	2	{(1,1,2)}	{(1,5,2)}	{(1,1,1),(2,3,2),(1,5,4)}
v_2	v_0	8	1	{(1,1,2),(1,5,1)}	{(1,5,2)}	{(1,1,1),(2,3,2),(1,5,1)}

Table 1: Changes at buffers at each interaction

- **Most recently selection model:** gives priority to the quantities that are generated more recently.
- **Proportional selection model:** the transferred quantity is selected proportionally, based on the origin.

r.s	r.d	r.t	r.q	$\mathbf{p}_{v_0}$	$\mathbf{p}_{v_1}$	$\mathbf{p}_{v_2}$
v_1	v_2	1	3	[0, 0, 0]	[0, 0, 0]	[0, 3, 0]
v_2	v_0	3	5	[0, 3, 2]	[0, 0, 0]	[0, 0, 0]
v_0	v_1	4	3	[0, 1.2, 0.8]	[0, 1.8, 1.2]	[0, 0, 0]
v_1	v_2	5	7	[0, 1.2, 0.8]	[0, 0, 0]	[0, 5.8, 1.2]
v_2	v_1	7	2	[0, 1.2, 0.8]	[0, 1.66, 0.34]	[0, 4.14, 0.86]
v_2	v_0	8	1	[0, 2.03, 0.97]	[0, 1.66, 0.34]	[0, 3.31, 0.69]

Table 2: Changes at buffers (Proportional)

Experiments

Datasets

- Bitcoin: a network which users exchange money (generated by real data)
- Prosper Loans: users which borrow money to other users

Dataset	#nodes	#interactions	avg. q_i
Bitcoin	12M	45.5M	34.4
Prosper Loans	88K	3.08M	76

Table 3: Characteristics of datasets

Results

- Runtime and memory footprint comparison

Dataset	Least recently	Most recently	Proportional
Bitcoin	31.77	9.17	7.25
Prosper Loans	0.089	0.082	0.209

Table 4: Runtime (sec) for each proposed model

Dataset	Least recently	Most recently	Proportional
Bitcoin	534MB	535MB	4.83GB
Prosper Loans	36.8MB	36.8MB	3.5GB

Table 5: Memory capacity for each proposed model

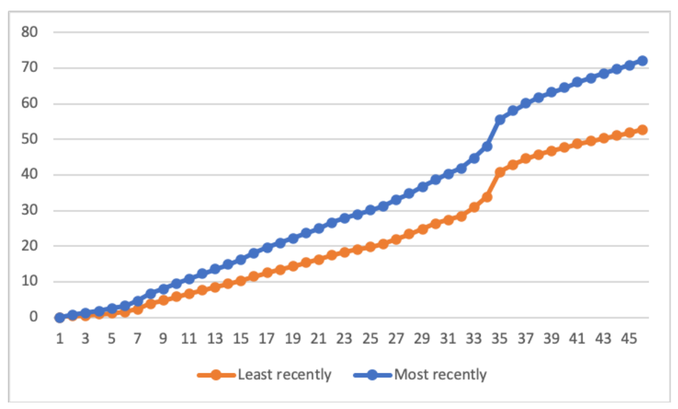


Figure 4: Time VS Interactions in Bitcoin

References

1) P. Buneman, S. Khanna and W.C Tan: Why and Where: A characterization of data provenance in Database Theory – ICDT, 2001

2) A. Chapman, H. V. Japadish, and P. Ramanam. Efficient provenance storage. In Proceedings of the ACM SIGMOD, 2008

3) G. Karvounarakis, Z. G. Ives, and V. Tannen. Querying data provenance. In Proceedings of the ACM SIGMOD, 2010

4) Rohit Kumar and Toon Calders. Information Propagation in Interaction Networks. In EDBT, 2017