TPII: tracking personally identifiable information via user behaviors in HTTP traffic

Yi LIU, Tian SONG, Lejian LIAO

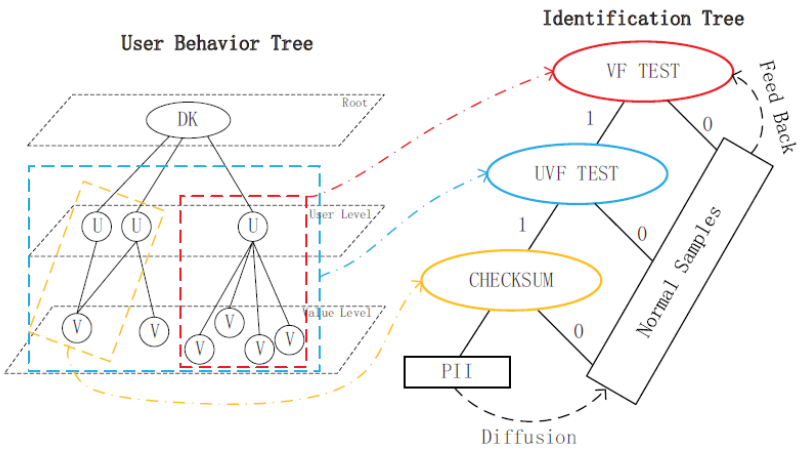
Frontiers of Computer Science, DOI: 10.1007/s11704-018-7451-z

Problems & Ideas

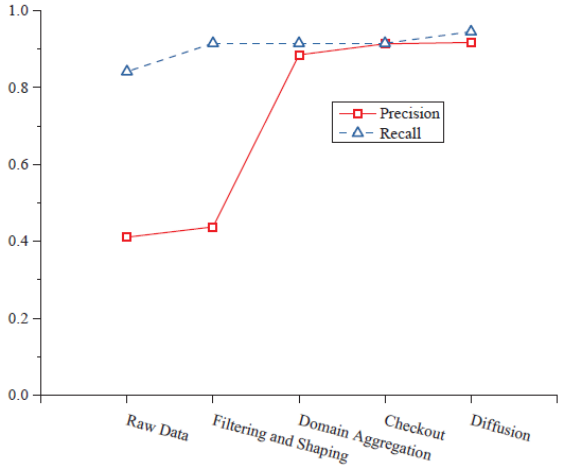
- Problems of accurately locate PIIs in the massive data of network traffic just like looking a needle in a haystack.
  - Where the PIIs is leak in the network traffic?
  - What kinds of PII?
- Ideas: User behaviors model
  - The massive traffic was changed a dataset with consists of five dimensions including user, domain, key, value, frequency. The five dimensions can represent user's cyber space behavior, that user accesses the network application service as domain, the value transmitted at the location as key, and transmits the frequency as F.
  - we can create user behavior tree model with the same domain-key
  - TPII applies three nave natures of PIIs to user behavior tree.

Main Contributions

- TPII algorithm



- Precision and recall of TPII approach with five process



- Comparison between our proposed TPII approach and baseline method.

PII Type	Baseline No. of DKs	Baseline FP	TPII No. of DKs	TPII FP	Comparison Common DKs	Comparison Unique to TPII	Comparison Unique to Baseline
IMEI	1182	33.93%	564	1.77%	514	50	1132
MAC	1345	70.78%	553	7.05%	493	60	1285
IDFA	510	30.20%	244	11.07%	49	195	315
Device ID	2578	48.95%	1585	27.38%	455	1130	1448
User ID	3438	70.78%	1621	19.43%	738	883	2555
Name	5484	83.40%	1898	35.35%	315	1583	3901
E-mail	528	76.14%	177	1.69%	18	159	369
Password	170	30.59%	60	8.33%	45	15	155
Phone	1383	82.50%	145	6.21%	92	53	1330
Location	3518	54.55%	221	7.24%	136	85	3433
Total	20136	58.18%	7382	12.55%	2855	4213	15923

- The TPII approach deal with 122k files in parallel process, and the parallel computing time of 213 million records reduced from 9 days, 9 hours and 17 minutes to 1 hour.