Towards Scalable and Dynamic Social Sensing in CPS Using Distributed Computing Framework (Award #: CNS-1566465)

PI: Dr. Dong Wang
Department of Computer Science & Engineering
University of Notre Dame

Abstract

With the rapid growth of online social media and ubiquitous Internet connectivity, social sensing has emerged as a new CPS application paradigm of collecting observations (often called claims) about the physical environment from humans or devices on their behalf. A fundamental problem in social sensing applications lies in effectively ascertaining the correctness of claims and the reliability of data sources without knowing either of them a priori, which is referred to as ***truth discovery***.

In this work, we propose a Scalable Streaming Truth Discovery Scheme (SSTD) to address three fundamental challenges in social sensing truth discovery process.

Introduction

Motivation – Who is telling the truth?

In the big data era, it's important to identify trustworthy information from an influx of noisy data contributed by unvetted sources from online social media.

EXAMPLE TWEETS ON CONTRADICTING CLAIMS IN OSU CAMPUS ATTACK, NOVEMBER, 2016

Tweet	Timestamp
OSU POSSIBLE SHOOTING: I am on campus near @OSUengineering TONS of police.	28 Nov 2016, 7:23 AM
There was a shooting at Ohio state please pray for people's safety #osu	28 Nov 2016, 7:47 AM EST
Liberals putting out fake claims about the terrorist attack. 1st not a shooting, 2nd not an American, 3rd not nazi but Islamic #osushooting	28 Nov 2016, 11:37 AM EST

Table 1. Example of Conflicting Claims

Three Fundamental Challenges

Dynamic Truth – when the true information dynamically changes (e.g. weather condition, football scores, stock price, etc.), how to effectively find such information in a timely manner?

Scalability – how to handle huge social sensing data streams in the era of Big Data?

Heterogeneity and Unpredictability – how to optimize the system when the data traffic dynamically changes and hard to predict?

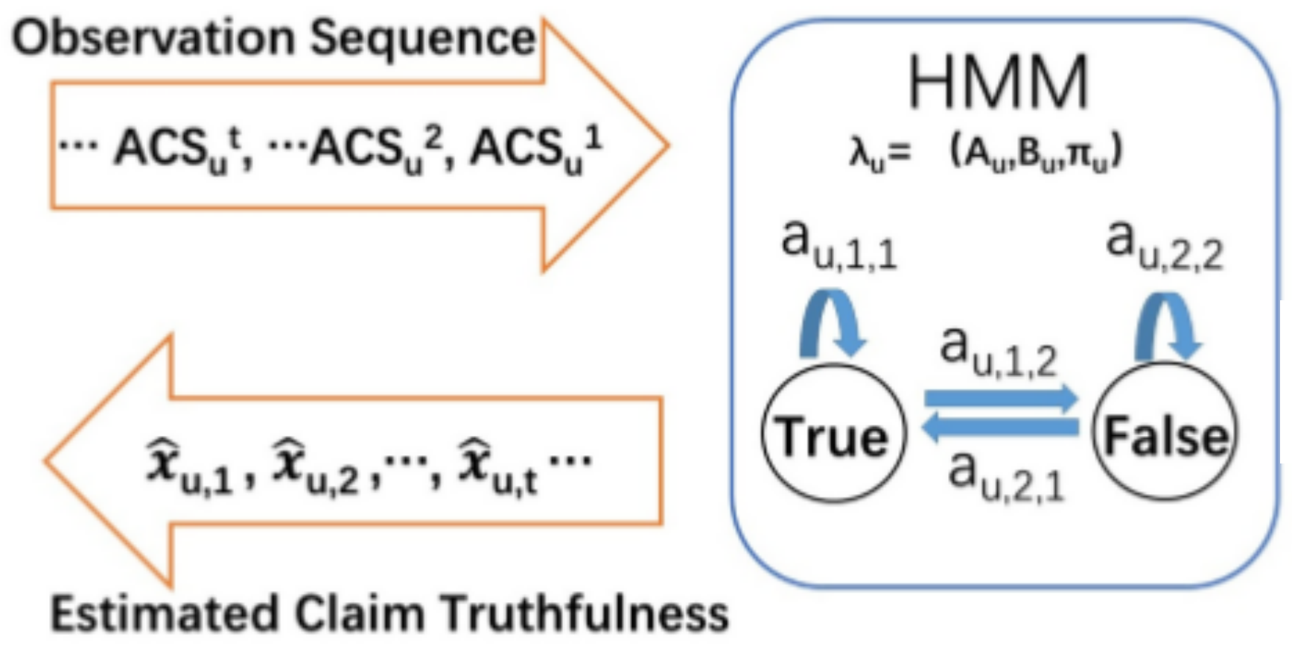
Problem Statement & Solution

We formulate a constrained optimization problem: Optimize the **efficiency** and **effectiveness** of the system given resource and deadline constraints.

$$\begin{aligned} &\text{maximize} \quad P(\hat{x}_{u,t} = x_{u,t} | S, C, R), \forall t > 0 \\ &\text{and} \quad P(w_j^{\Delta t} \le dl_j), \forall 1 \le j \le J \\ &\text{s.t. } RC^k \quad \text{is satisfied} \quad \forall 1 \le k \le K \end{aligned}$$

S_i	The ith source
C_u	The uth claim
$R_{i,u}^t$	The report made by S_i about the true value of C_u at time t
$\hat{x}_{u,t}$	The estimation of the true value of C_u at time t
$x_{u,t}$	The ground truth label of the claim C_u at time t
$w_j^{\Delta t}$	Worse Case Execution Time for processing $D_j^{\Delta t}$
dl	A set of soft deadlines for the TD jobs
RC^k	A set of resource constrains that defines the maximum available resources for k-th node

- **Dynamic Truth**
  - Designed a Hidden Markov Model (HMM) based streaming algorithm to explicitly model the transition of truth.
  - We model source quality based on semantic links (namely Contribution Score) between each source and claim. The weight of each link reflects both the extent that a source believes a claim is true and the authority of the source.
  - We model the crowd opinion (ACS) at a specific time instance by aggregating Contribution Scores. Intuitively, this "observed" crowd opinion is related to the "hidden" state of truth.


Observation Sequence
··· ACS$_u^t$, ···ACS$_u^2$, ACS$_u^1$
HMM
λ$_u$= (A$_u$,B$_u$,π$_u$)
a$_{u,1,1}$
a$_{u,2,2}$
a$_{u,1,2}$
a$_{u,2,1}$
True
False
$\hat{x}_{u,1}$, $\hat{x}_{u,2}$,···, $\hat{x}_{u,t}$···
Estimated Claim Truthfulness

Attitude Score $(\rho_{i,u}^t)$

Uncertainty Score $(\kappa_{i,u}^t)$

Independent Score: $(\eta_{i,u}^t)$

$$CS_{i,u}^t = \rho_{i,u}^t \times (1 - \kappa_{i,u}^t) \times \eta_{i,u}^t$$

$$ACS_u^t = \sum_{t-sw}^{t} CS_{i,u}^t$$

Figure 1. HMM based Dynamic Truth Discovery Scheme

- **Scalability**
  - Distributed system implementation using HTCondor and Work Queue
- **Data Heterogeneity and Unpredictability**
  - Designed a real-time control system that dynamically optimizes the system based on incoming data and system performance feedback
  - Deadline driven – the system tries its best to meet the deadline requirements
  - Flexible control knobs allowing fine-grained system performance tuning

Contact

Dr. Dong Wang
Email: dwang5@nd.edu
Social Sensing and Collective Computing Lab: *http://www3.nd.edu/~sslab/*
Department of Computer Science & Engineering, University of Notre Dame

System Implementation

Dynamic Task Manager (DTM) – master process of Work Queue. It initializes a Worker Pool and dynamically spawns new tasks into the Task Pool.
TD Job – each TD job is an instance of our HMM based truth discovery algorithm
Local Control Knob (LCK) – priority assignment for each TD job, a job with higher priority potentially runs faster
Global Control Knob (GLK) – the global resources assigned to the system
Feedback Control System (FCS) – A PID feedback control loop is implemented to monitor the execution time of each job. It dynamically tunes both GCK and LCK to optimize system performance.

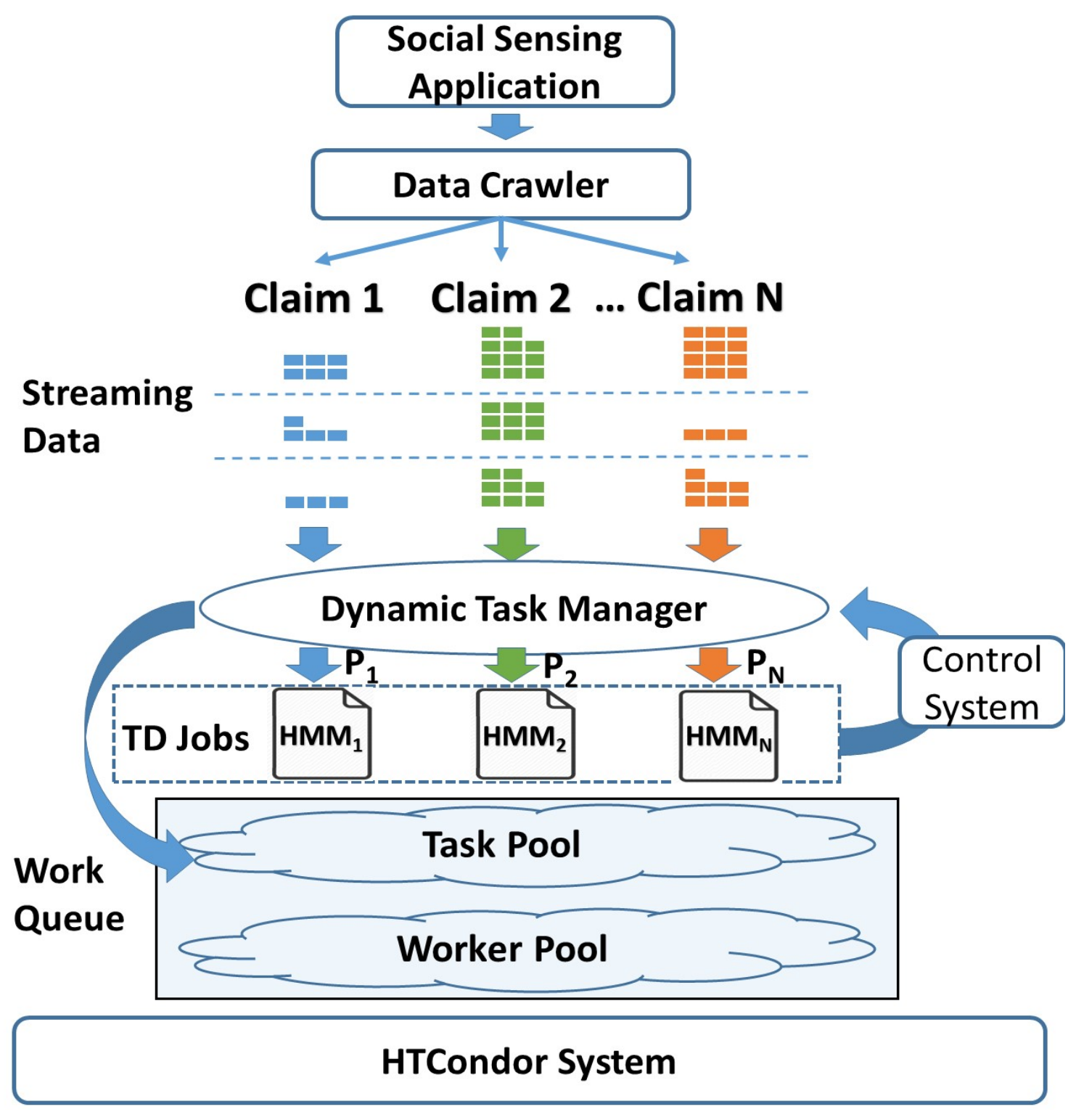


Figure 2. SSTD System Architecture

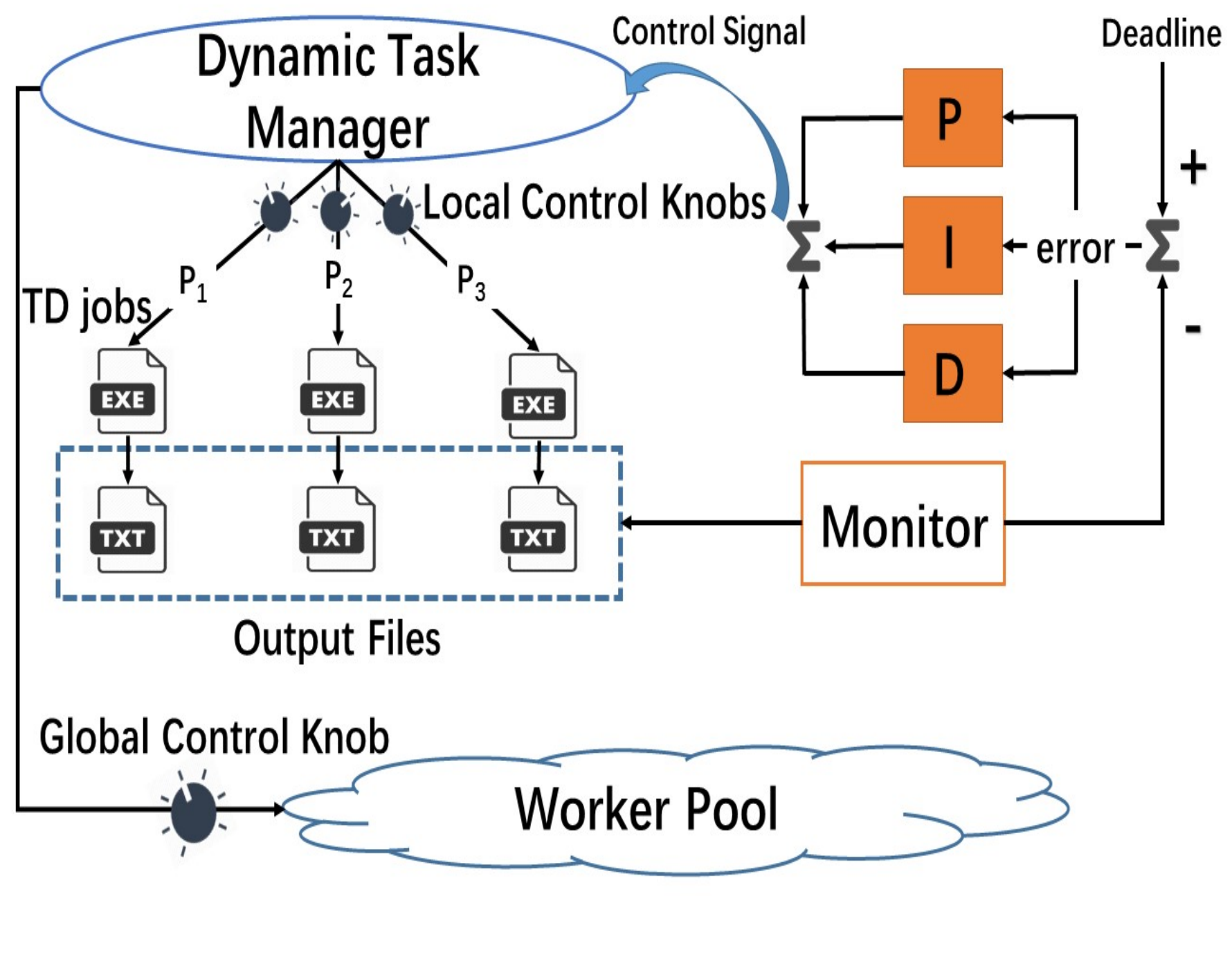


Figure 3. Dynamic Feedback Control System

Experiment Results

Data Trace	Paris (Charlie Hebdo) Shooting	Boston Bombing	College Football
Start Date	Jan. 1 2015	Apr. 15 2013	Sep. 30 2016
Time Duration	3 days	4 days	3 days
Search Keywords	Paris, Shooting, Charlie Hebdo	Bombing, Marathon, Attack	Team/College names
# of Reports	253,798	553,609	429,019
# of Sources	217,718	493,855	413,782

Table 2. Data Trace Summary

We evaluate our SSTD system based on both effectiveness and efficiency. The results show that our scheme has achieved significant performance gains in terms of classification accuracy, scalability and the ability to meet deadline requirements.

TRUTH DISCOVERY RESULTS - COLLEGE FOOTBALL

Method	Accuracy	Precision	Recall	F1-Score
SSTD	**0.801**	**0.661**	**0.792**	**0.723**
DynaTD	0.765	0.471	0.570	0.515
TruthFinder	0.612	0.542	0.455	0.495
RTD	0.752	0.555	0.649	0.598
CATD	0.736	0.542	0.764	0.634
Invest	0.722	0.478	0.716	0.574
3-Estimates	0.674	0.396	0.677	0.501

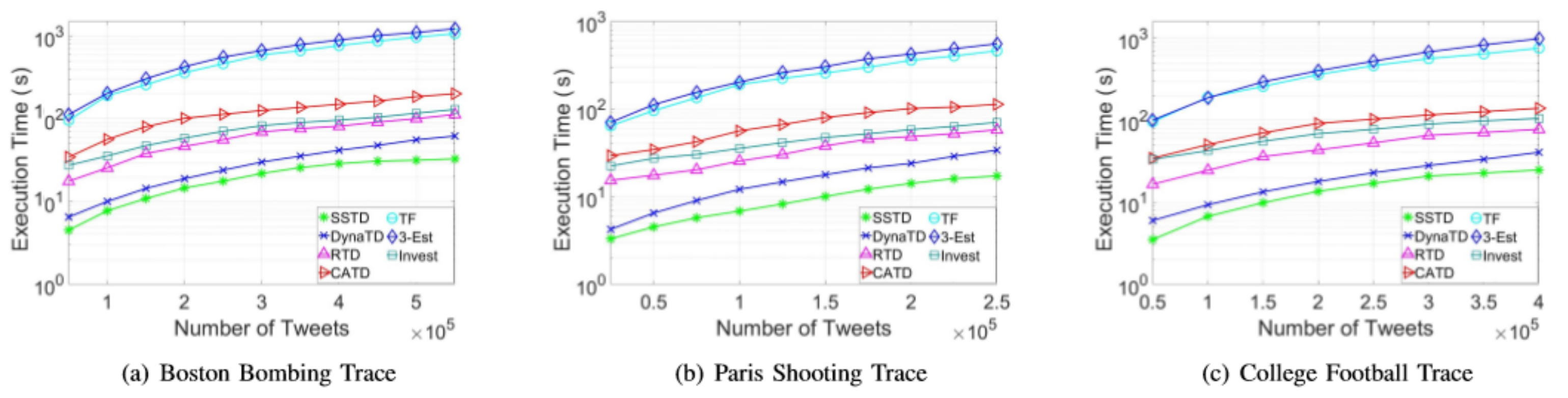
TRUTH DISCOVERY RESULTS - PARIS SHOOTING

Method	Accuracy	Precision	Recall	F1-Score
SSTD	**0.802**	**0.834**	**0.905**	**0.872**
DynaTD	0.731	0.822	0.788	0.805
TruthFinder	0.616	0.653	0.806	0.721
RTD	0.753	0.791	0.823	0.807
CATD	0.669	0.689	0.760	0.723
Invest	0.661	0.722	0.780	0.750
3-Estimates	0.647	0.704	0.765	0.733

TRUTH DISCOVERY RESULTS - BOSTON BOMBING

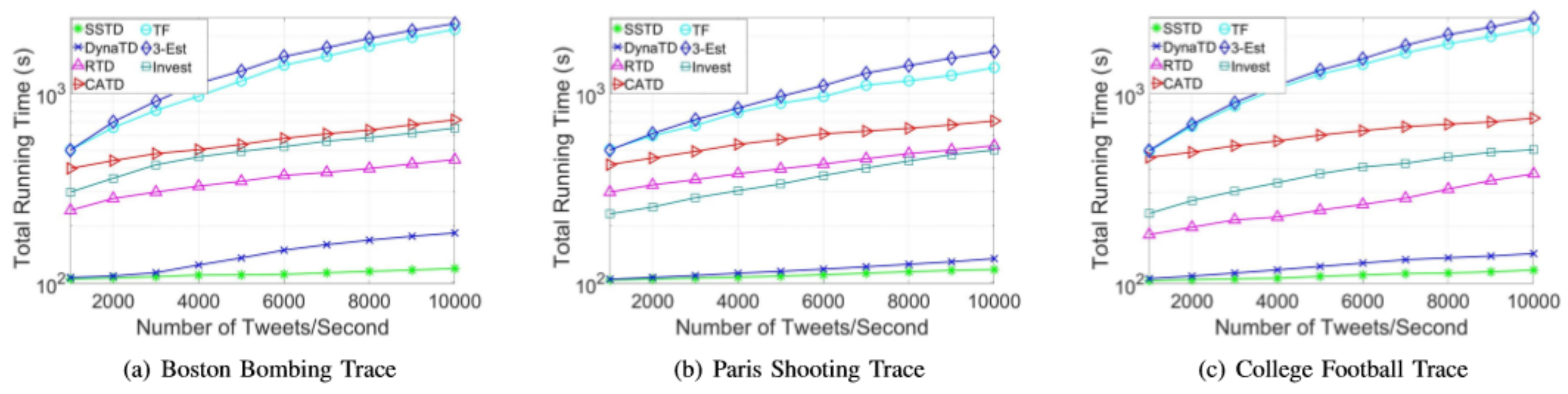
Method	Accuracy	Precision	Recall	F1-Score
SSTD	**0.828**	**0.834**	**0.831**	**0.833**
DynaTD	0.722	0.811	0.756	0.783
TruthFinder	0.653	0.689	0.787	0.734
RTD	0.763	0.748	0.824	0.784
CATD	0.667	0.764	0.748	0.751
Invest	0.609	0.639	0.626	0.632
3-Estimates	0.616	0.626	0.807	0.705

Table 3. Evaluation on Classification Accuracy



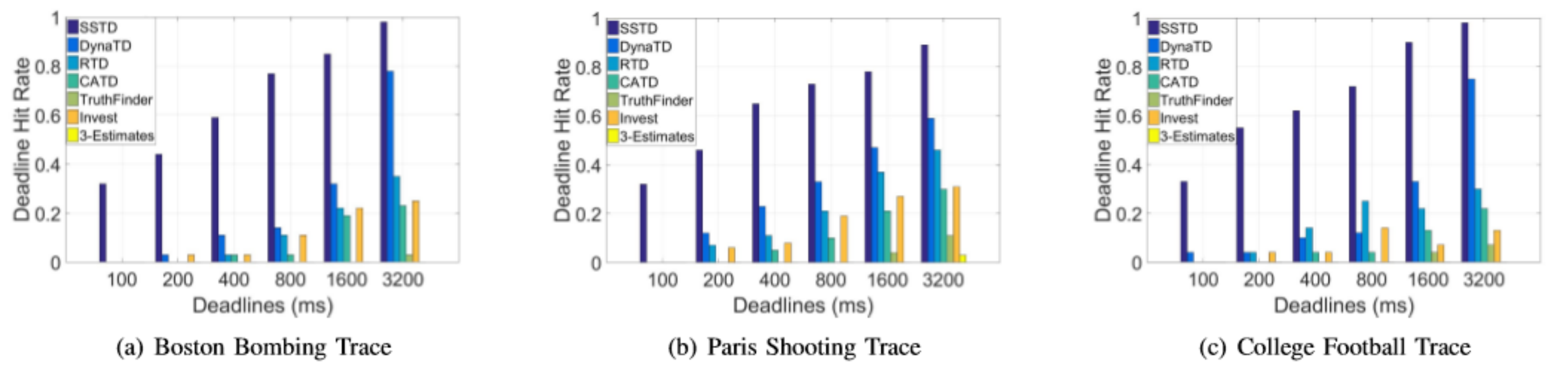
(a) Boston Bombing Trace (b) Paris Shooting Trace (c) College Football Trace

Figure 4. Scalability Analysis



(a) Boston Bombing Trace (b) Paris Shooting Trace (c) College Football Trace

Figure 5. Total Running Time vs. # of tweets per Sec



(a) Boston Bombing Trace (b) Paris Shooting Trace (c) College Football Trace

Figure 6. Deadline Hitting Rates

Publication

- Daniel Zhang, Chao Zhang, Dong Wang, Doug Thain, Xin Mu, Greg Madey and Chao Huang. Towards Scalable and Dynamic Social Sensing Using A Distributed Computing Framework, The 37th IEEE International Conference on Distributed Computing (ICDCS 2017)
- Daniel Zhang, Dong Wang, Yang Zhang. Constraint-Aware Dynamic Truth Discovery in Big Data Social Media Sensing, 2017 IEEE International Conference on Big Data (IEEE BigData 2017)
- Chao Huang, Dong Wang, Nitesh Chawla. Scalable Uncertainty-Aware Truth Discovery in Big Data Social Sensing Applications for Cyber-Physical Systems, IEEE Transactions on Big Data