FlowCog: Context-aware Semantics Extraction and Analysis of Information Flow Leaks in Android Apps

Xiang Pan, Yinzhi Cao, Xuechao Du, Boyuan He, Gan Fang, Yan Chen.

Northwestern University, Johns Hopkins University
Zhejiang University, Google

Roadmap

1. **Motivating Example**
2. FlowCog Overview
3. Design
    a. View Dependency Explorer
    b. Flow and Semantics Correlation Inference
4. Implementation
5. Evaluation & Case Study
6. Conclusion

Main_activity

<Button an

"*perfomUp*

Pressure: 1012.0 hPa
Humidity: 51 %
Sunrise: 1:14 PM

MainActivity:p

(2) task=new Up ... (userLocation, **allowShareLoc**.isChecked())

task.execute

Wind: 1.4 m/s ↗

UpdateWeath ... ound()

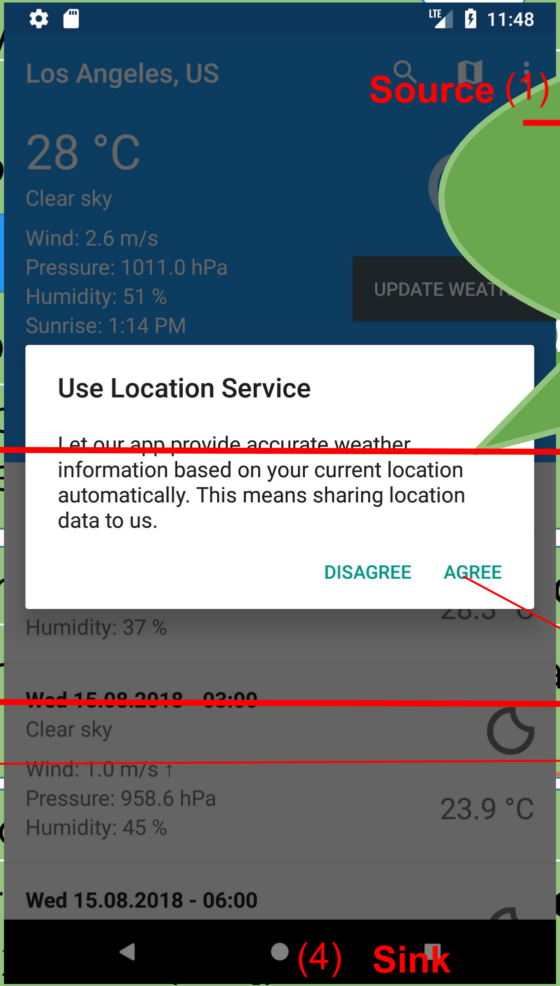
(3) **if** (allowShar ... ata(**location**)

else showDi

ServerApi.p

req = prepar ... **ocation**)

HttpClient.e



11:48

Los Angeles, US

28 °C

Clear sky

Wind: 2.6 m/s
Pressure: 1011.0 hPa
Humidity: 51 %
Sunrise: 1:14 PM

UPDATE WEAT

Use Location Service

Let our app provide accurate weather information based on your current location automatically. This means sharing location data to us.

DISAGREE AGREE

Humidity: 37 %

Wed 15.08.2018 - 03:00

Clear sky

Wind: 1.0 m/s ↑
Pressure: 958.6 hPa
Humidity: 45 %

23.9 °C

Wed 15.08.2018 - 06:00

Let our app provide accurate weather information based on your current location, this ...

“Share location to automatically update city”

PosBtnListener:onClick()

(3) ServerApi.postData(location)

Cannot tell which flow is legitimate!

[MockDroid] B l, Alastair R., et al. "Mockdroid: tradi for application fun rtphones." Proceedings of the 12th workshop on mobile computing systems a ACM, 2011.

[SmartDroid] Zheng, Cong, et al. " Proceedings of the second ACM

[AppIntent] Yang, Zhemin f the 2013 ACM SIGSAC confer

[Epicc] Octeau, Damien, et al. "Effectiv appl icc: An esse rds holistic security analysis." Effective Inter-Component Commun Mapping in An c: An Essent ds Holistic Security Ana (2013).

[AmanDroid] Wei, Fengguo, Sankardas Roy, and Xinming O android: A precise and gene al inter-component data flow analysis framework for security vetting of android apps." Proceedings of the 2014 ACM SIGS Conference on Computer and Communications Security. ACM, 2014.

[TaintDroid] Enck, William, et al. "TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones." AC M Transactions on Computer Systems (TOCS) 32.2 (2014): 5.

[Flow Arzt, Steven, Flowdro e context, flow, field, object-sensitive and ware ta fo gplan Notices4 259

android apps." Pr Engi

[A n: Revealing data flows in android applic d Applications (AINA), 2015 IE 015.

[Droidsafe] Gordon, Micha et al. "Infor on Flow Analysis of Android Applications in DroidSafe." NDSS. V 15. 2015.

[VetDroid] Pravin, Ms Nigan Paridhi. "Vetdroid: Analysis using permission for vetting undesirable behaviours in android applications." International Journal of Innovative and Emerging Research in Engineering 2.3 (2015): 131-136.

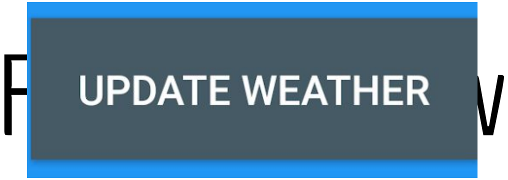
58% flows are legitimate!

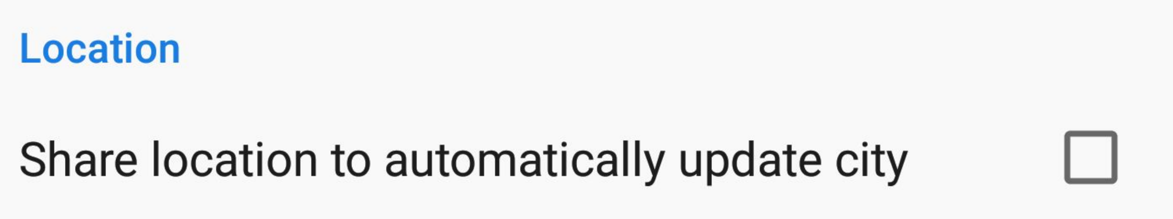
Big burden on users!

Roadmap

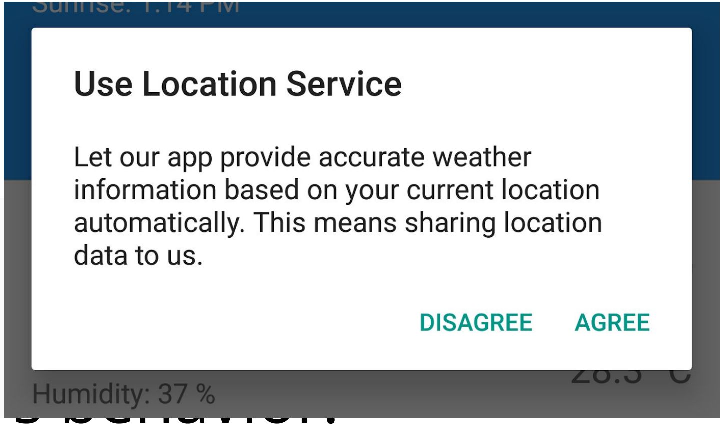
1. Motivating Example
2. **FlowCog Overview**
3. Design
   a. View Dependency Explorer
   b. Flow and Semantics Correlation Inference
4. Implementation
5. Evaluation & Case Study
6. Conclusion

F view




Location

Share location to automatically update city □


Sunrise: 1:14 PM

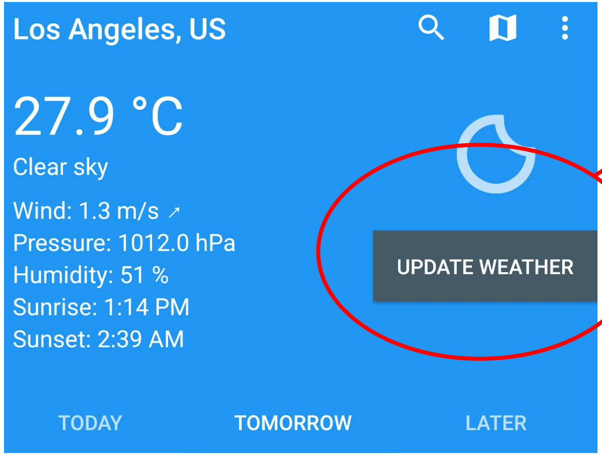
Use Location Service

Let our app provide accurate weather information based on your current location automatically. This means sharing location data to us.

Humidity: 37 %

- Associate the semantics with flow's behavior?

- High level steps:
  - Associate each flow with its related views via static analysis and an optional dynamic analysis.
  - Extract view semantics. (e.g., “Update Weather”)
  - Determine if semantics provides information about flow behavior.

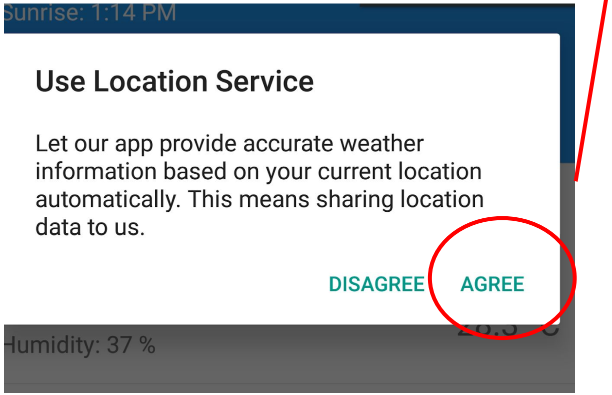


Los Angeles, US

27.9 °C

Clear sky

Wind: 1.3 m/s ↗
Pressure: 1012.0 hPa
Humidity: 51 %
Sunrise: 1:14 PM
Sunset: 2:39 AM



Sunrise: 1:14 PM

Use Location Service

Let our app provide accurate weather information based on your current location automatically. This means sharing location data to us.

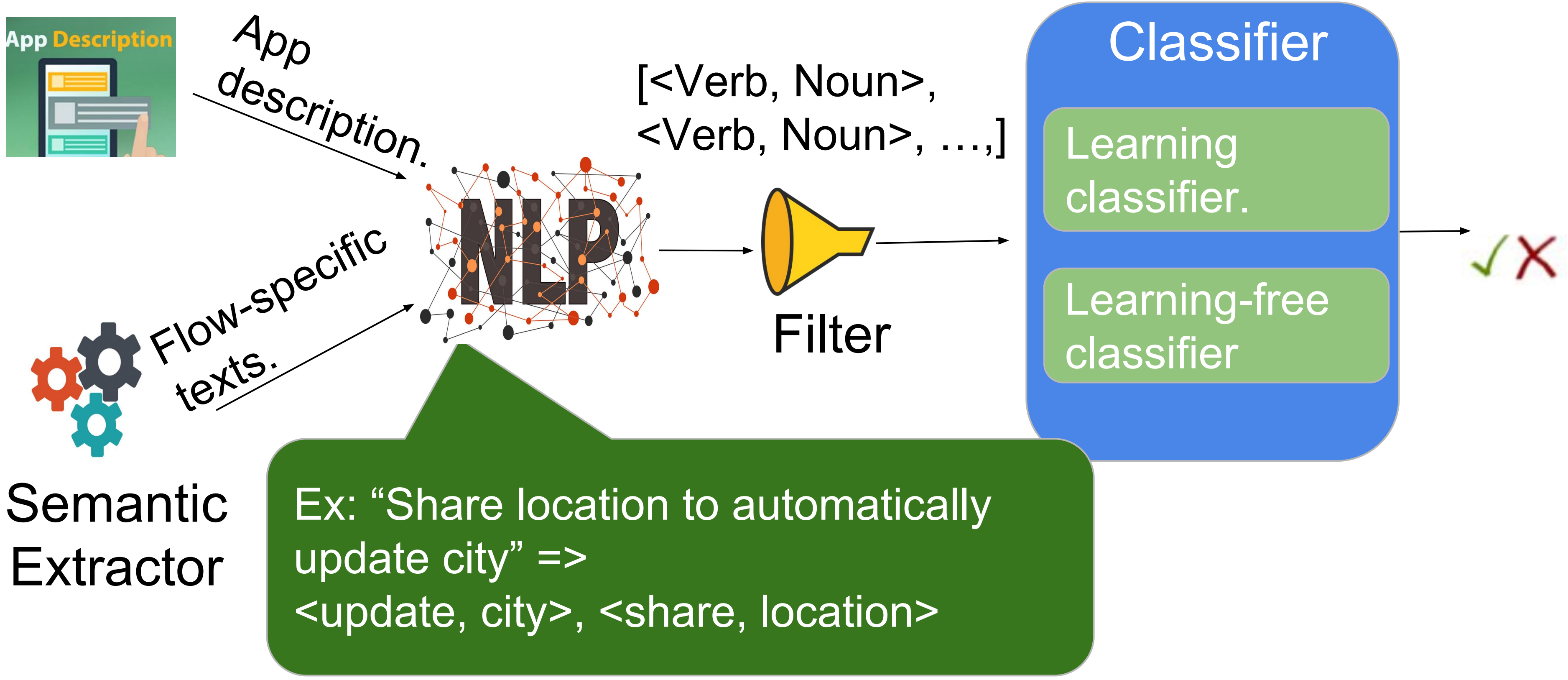
Humidity: 37 %

Activation Event.

```
UpdateWeatherAsyncTask:doInBackground
  if (allowShareLoc) sendData
  else ...
```

Guarding Condition

FlowCog: Flow and Semantics Correlation Inference (2/2)



App description.

Flow-specific texts.

Semantic Extractor

[<Verb, Noun>, <Verb, Noun>, …,]

Filter

Classifier

Learning classifier.

Learning-free classifier

✓ ✗

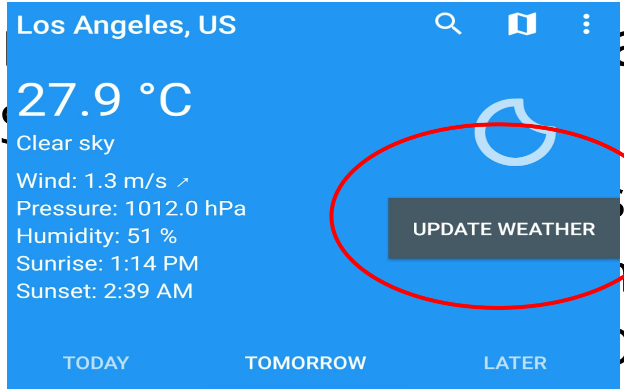
Ex: “Share location to automatically update city” =>
<update, city>, <share, location>

Roadmap

1. Motivating Example
2. FlowCog Overview
3. **Design**
    1. View Dependency Explorer
    2. Flow and Semantics Correlation Inference
4. Implementation
5. Evaluation & Case Study
6. Conclusion

Design: View Dependency Explorer

- For alysis p
  - S statemer at creat s) Dialog cla

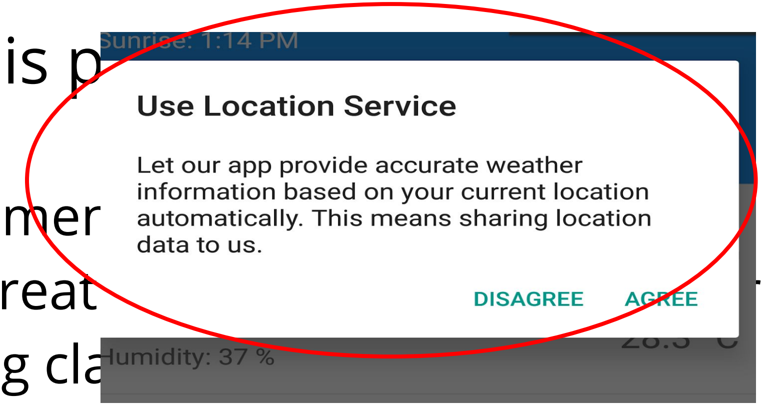


Los Angeles, US

27.9 °C

Clear sky

Wind: 1.3 m/s ↗
Pressure: 1012.0 hPa
Humidity: 51 %
Sunrise: 1:14 PM
Sunset: 2:39 AM



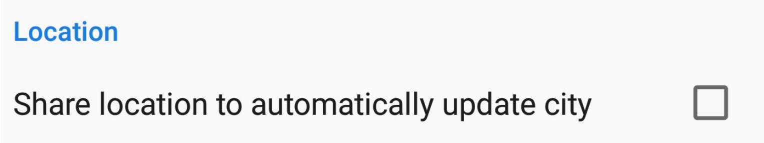
Sunrise: 1:14 PM

Use Location Service

Let our app provide accurate weather information based on your current location automatically. This means sharing location data to us.

Humidity: 37 %

- Sink:
  - Statements in given data flow.
  - Guarding condition statements.
  - All the activation events’ registration statements.



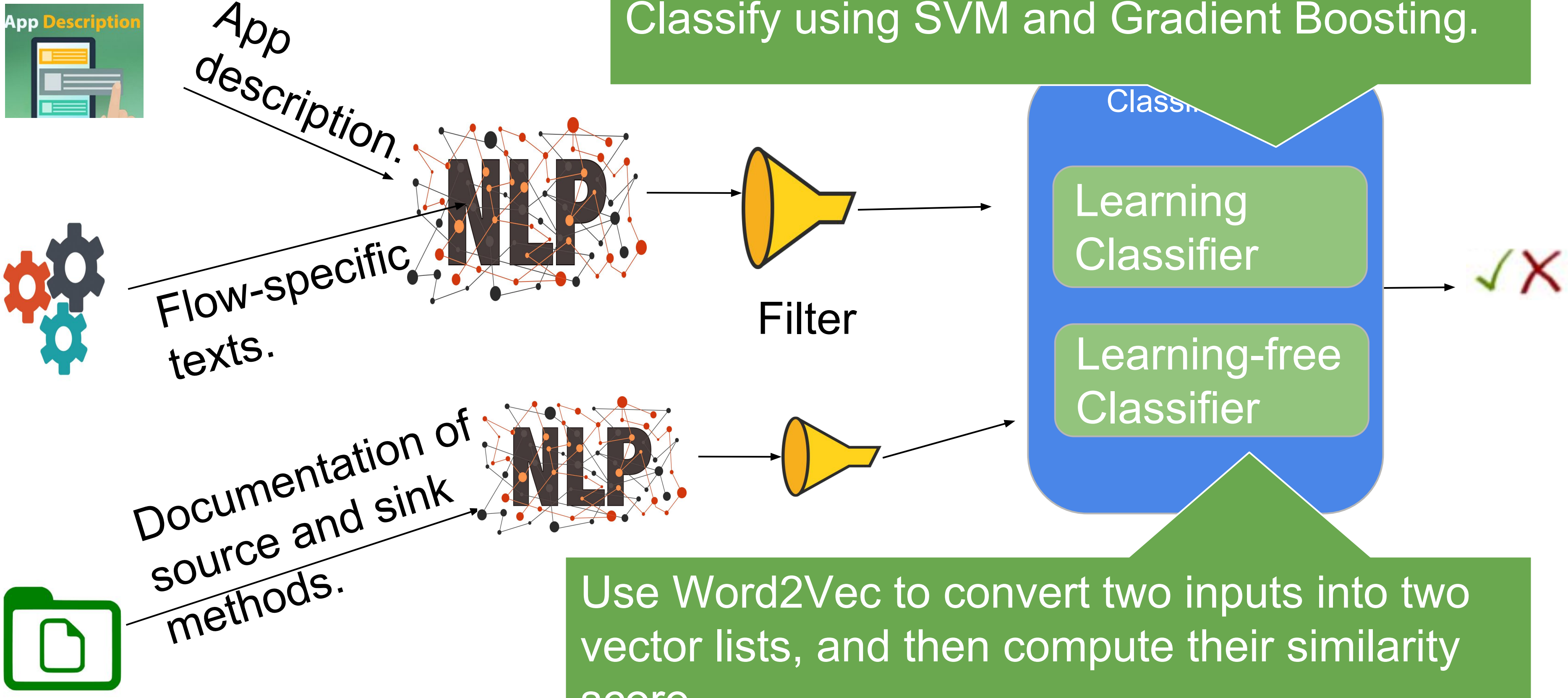
Location

Share location to automatically update city □

- Use IFDS framework provided by FlowDroid.

Design: Flow and Semant

Vectorize input using TF-IDF.
Classify using SVM and Gradient Boosting.



Use Word2Vec to convert two inputs into two vector lists, and then compute their similarity score.

Roadmap

1. Motivating Example
2. FlowCog Overview
3. Design
    a. View Dependency Explorer
    b. Flow and Semantics Correlation Inference
4. **Implementation**
5. Evaluation & Case Study
6. Conclusion

Implementation

Component	Language	Loc
Flow-related Semantics Extraction	Java	~12,000
Classifier	Python	~3,000
Dynamic Analysis	Python, Java	~1,000
Total	Python, Java	~16,000

Roadmap

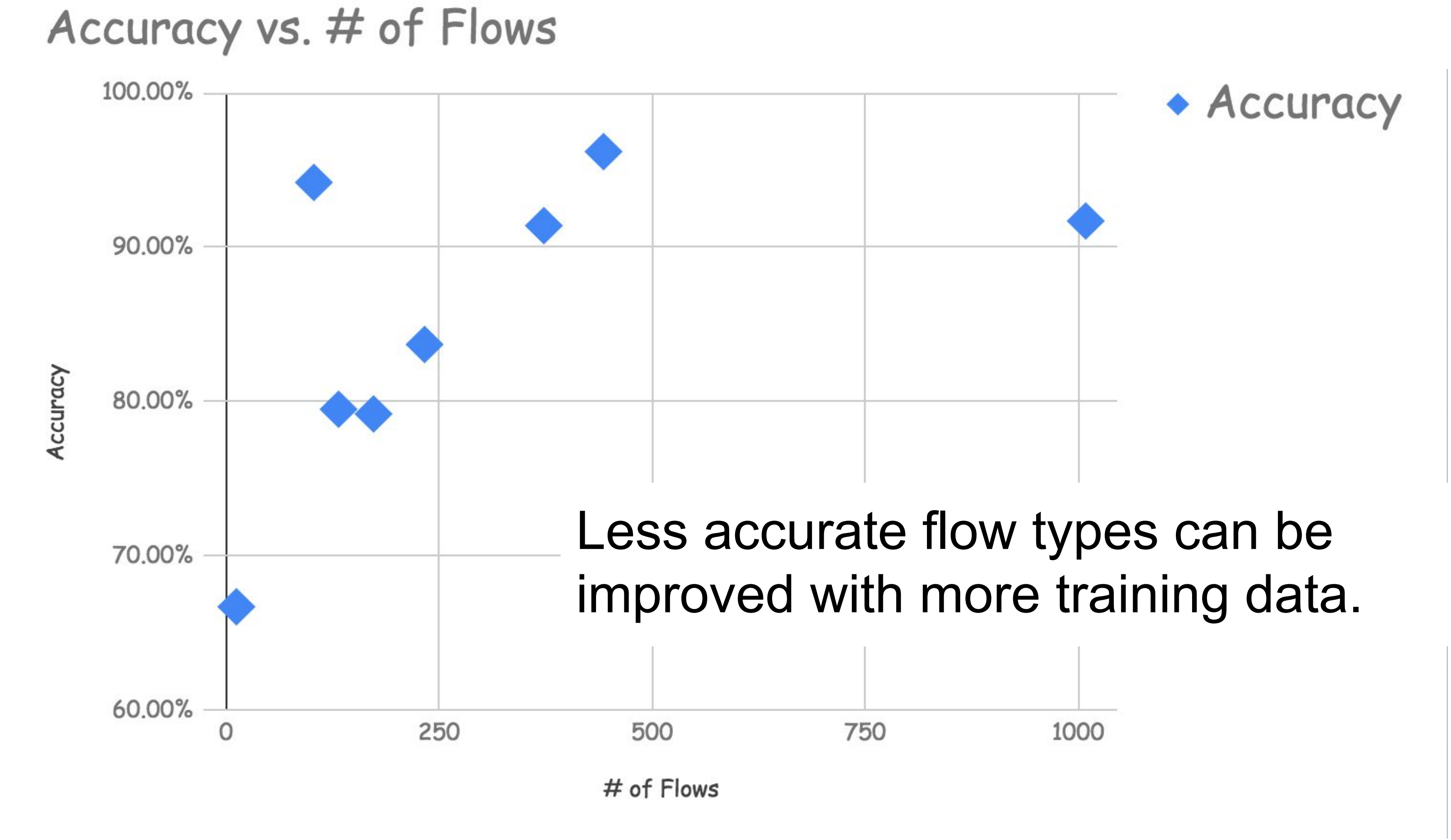
1. Motivating Example
2. FlowCog Overview
3. Design
   a. View Dependency Explorer
   b. Flow and Semantics Correlation Inference
4. Implementation
5. **Evaluation & Case Study**
6. Conclusion

Evaluation: Ground Truth

Type	**Apps**	**Apps with Flows**	**Legitimate Flows**	**Malicious Flows**	**Total Flows**
Benign	1,299/4,500	361	688	355	1,043
Malicious [Drebin dataset]	586/1,500	255	675	624	1,299
Overall	1,885/6,000	616	**1,363**	**979**	2,342

Evaluation: FlowCog Achieves High Accuracy.

Type	**Flows**	**Precision**	**Recall**	**Accuracy**
Benign	1,043	90.3%	95.1%	90.7%
Malicious	1,299	89.9%	91.0%	89.6%
Overall	2,342	90.1%	93.1%	90.2%

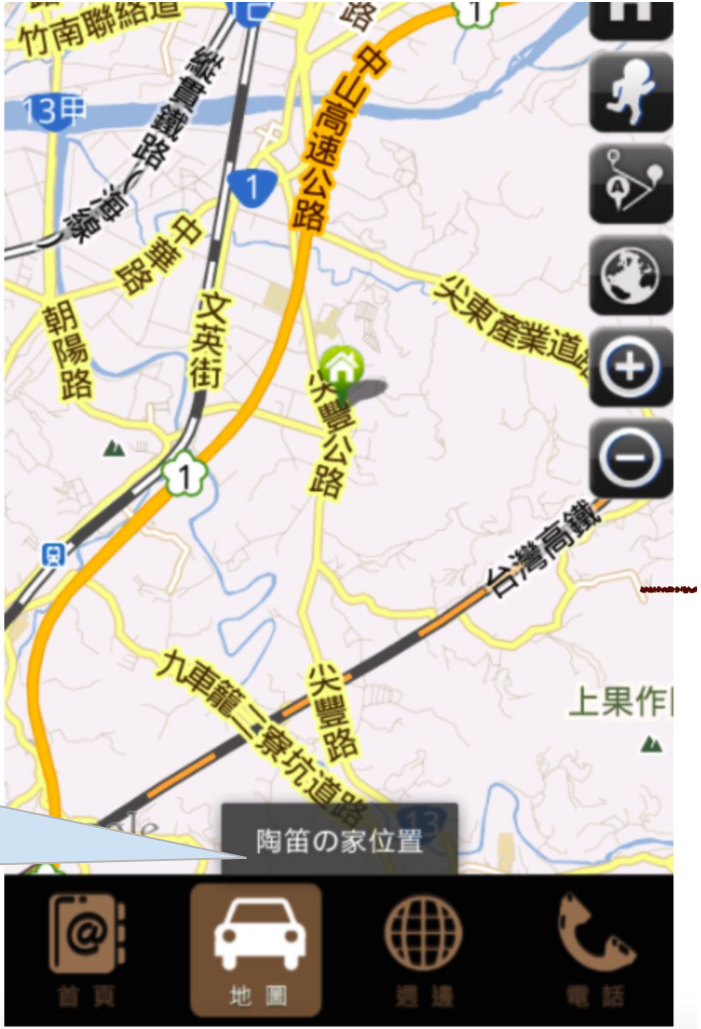


Less accurate flow types can be improved with more training data.

Case Study: Home of Ocarina

- Leaks out users’ geo-location.
- Labeled as legitimate because of extracted semantics.

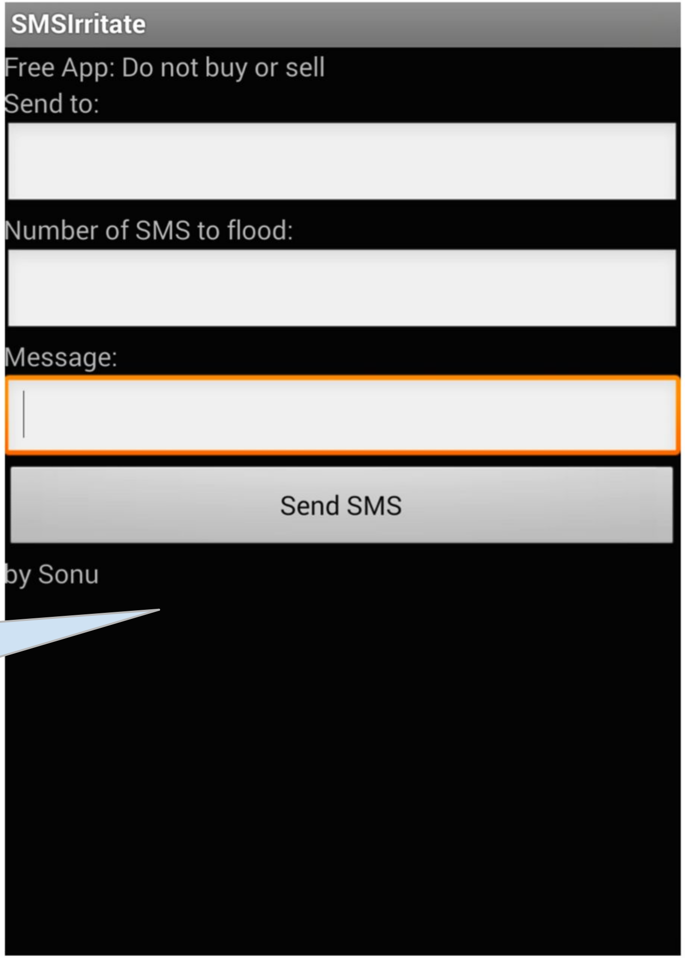
“Map”, “The location of home of Ocarina”



Case Study: SMS Irritate

- Leaks out user-specific information via SMS.
- Labeled as legitimate.

“Send SMS”, “Number of SMS to flood”, “Message”



SMSIrritate

Free App: Do not buy or sell

Send to:

Number of SMS to flood:

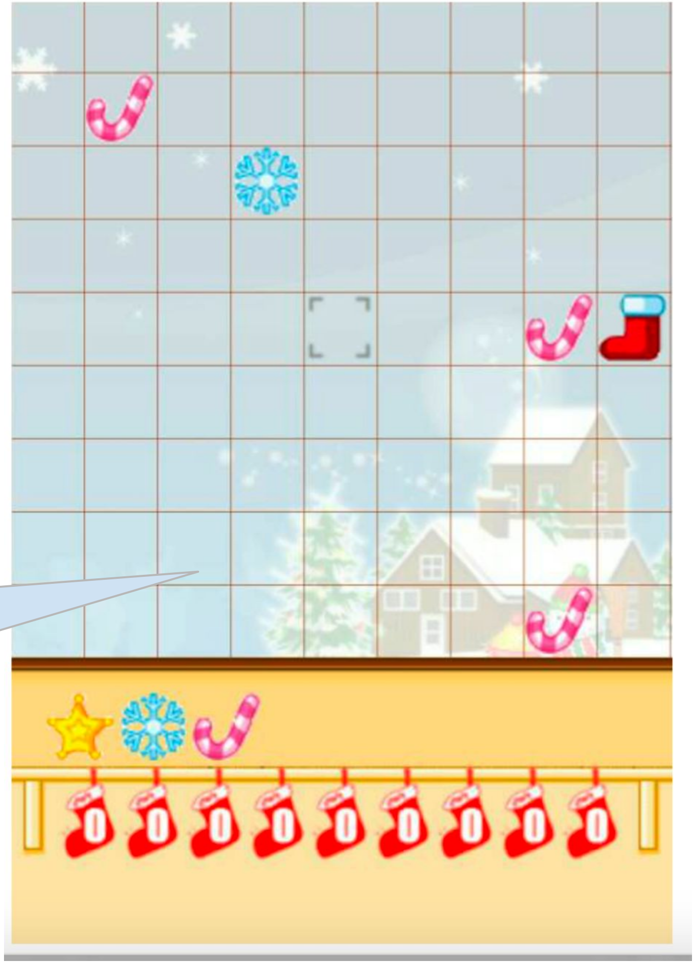
Message:

by Sonu

Case Study: Merry Christmas

- Leaks out users' information to Internet.
- Labeled as malicious.

"Move the box to the target empty position..."



Conclusion

- FlowCog is **the first system** to extract flow-specific semantics.
- FlowCog adopts NLP techniques to associate flow-specific semantics with flow behaviors.
- Our evaluation results show that FlowCog can achieve a precision of **90.1%** and a recall of **93.1%.**

Thanks!

FlowCog open-source at: **https: //github.com/SocietyMaster/FlowCog**