Into the Deep Web:
Understanding E-commerce Fraud from Autonomous Chat with Cybercriminals

Peng Wang, Xiaojing Liao, Yue Qin, XiaoFeng Wang
Indiana University Bloomington

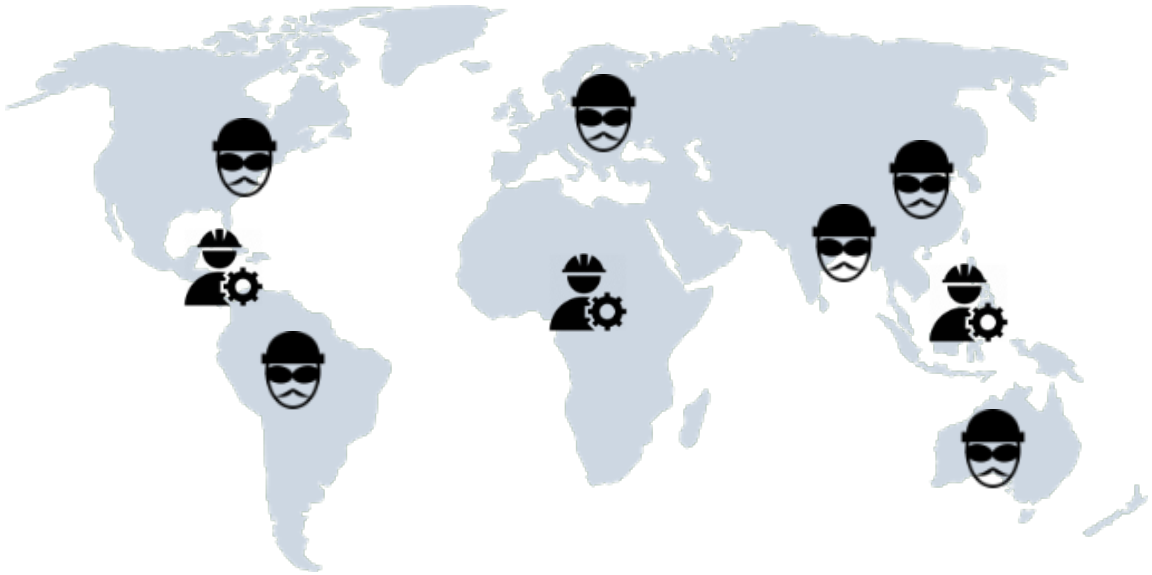
INFORMATICS, COMPUTING, AND ENGINEERING

E-commerce fraud

online
fraudsters

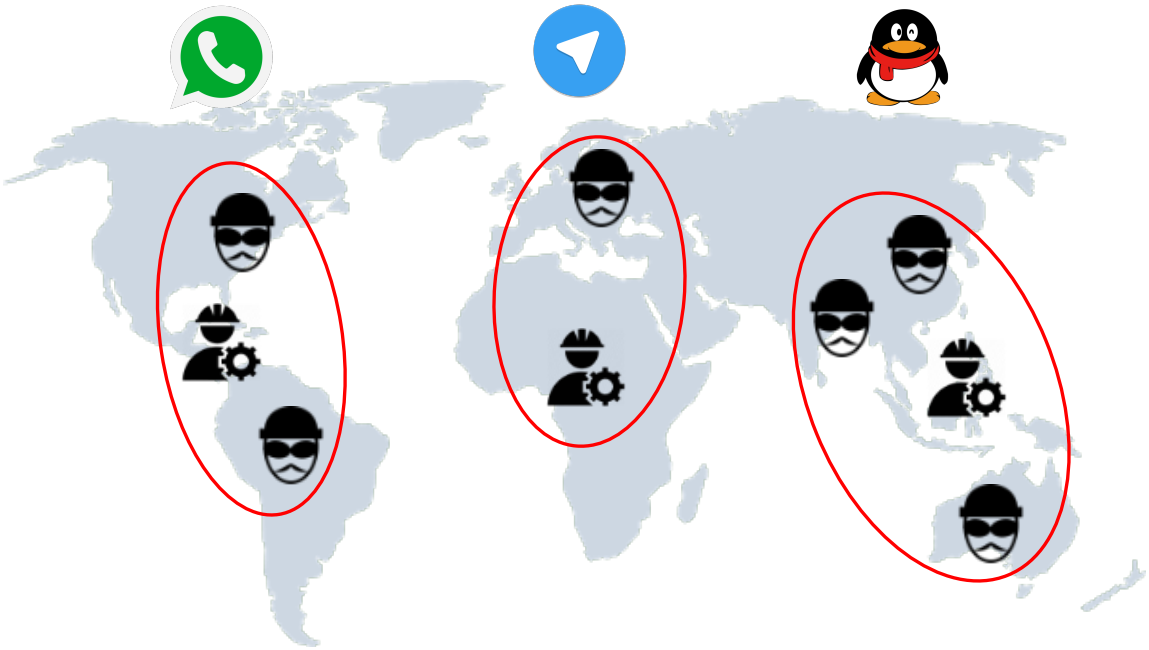
JD.COM

Crowdsourcing in e-commerce fraud



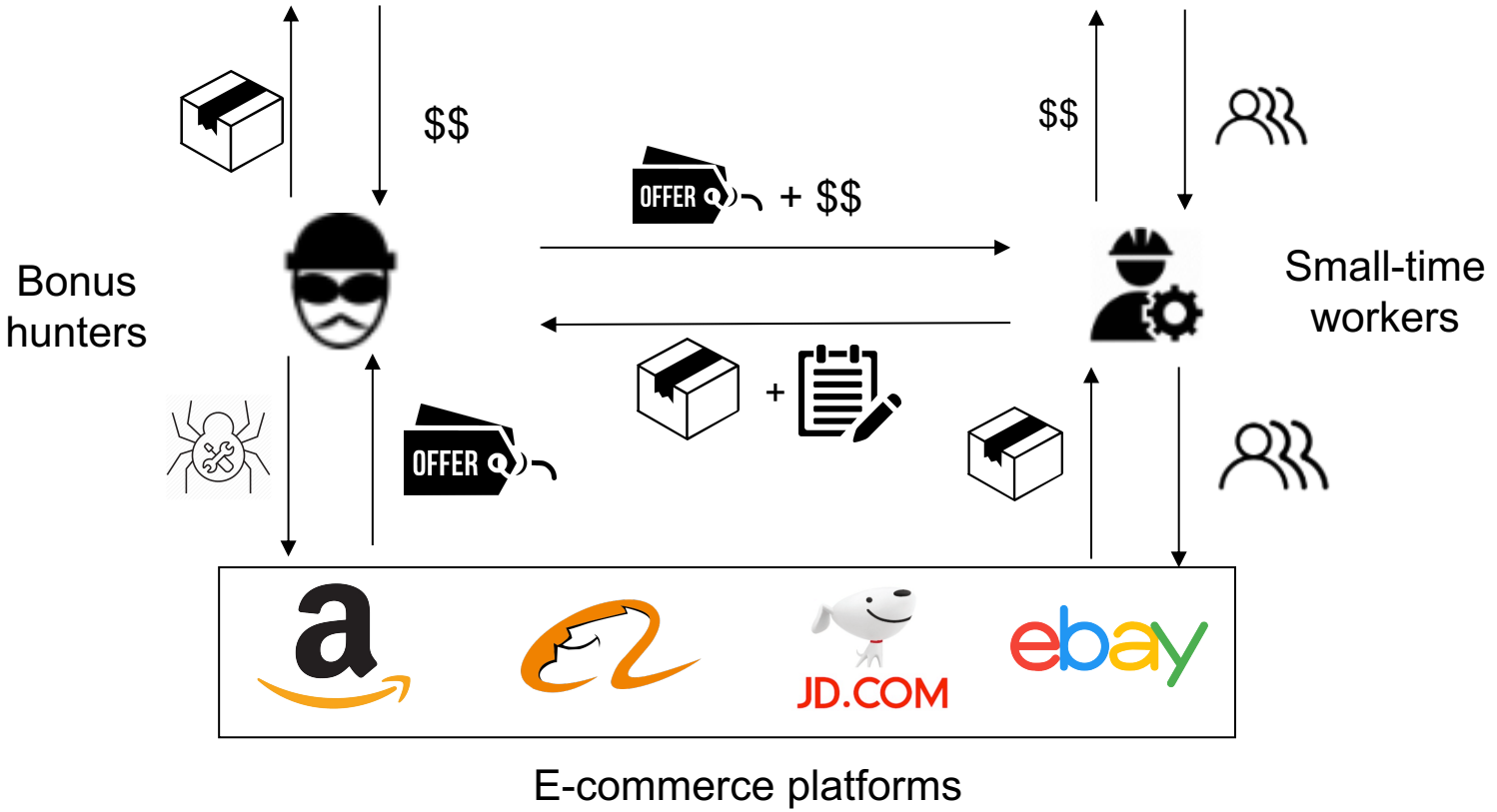
Crowdsourcing

Crowdsourcing via IM

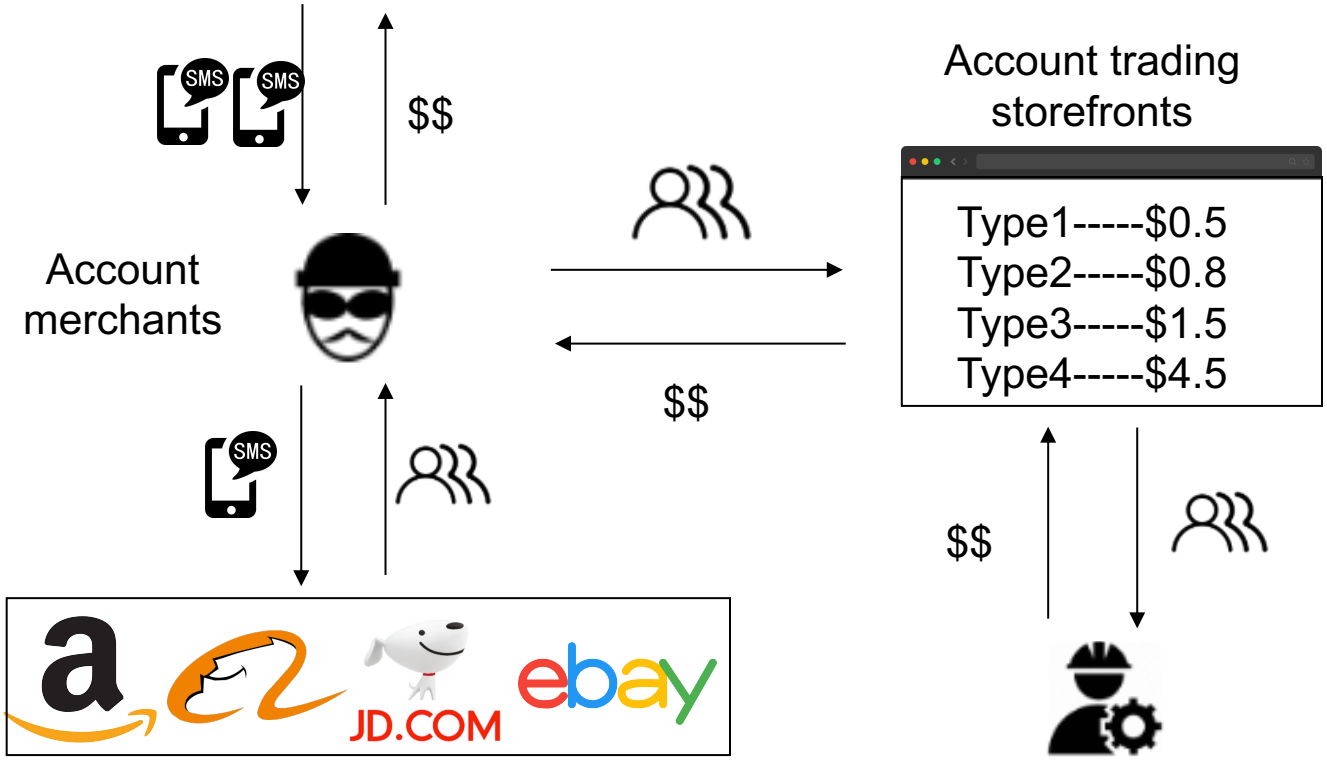


Crowdsourcing
via Instant Messaging (IM)

Bonus hunting

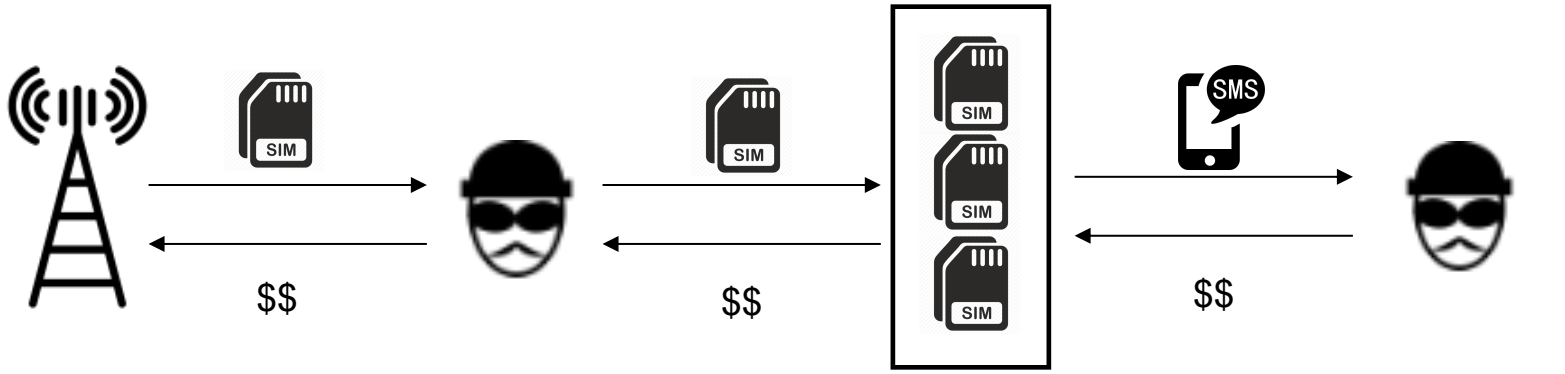


Fraud account trading

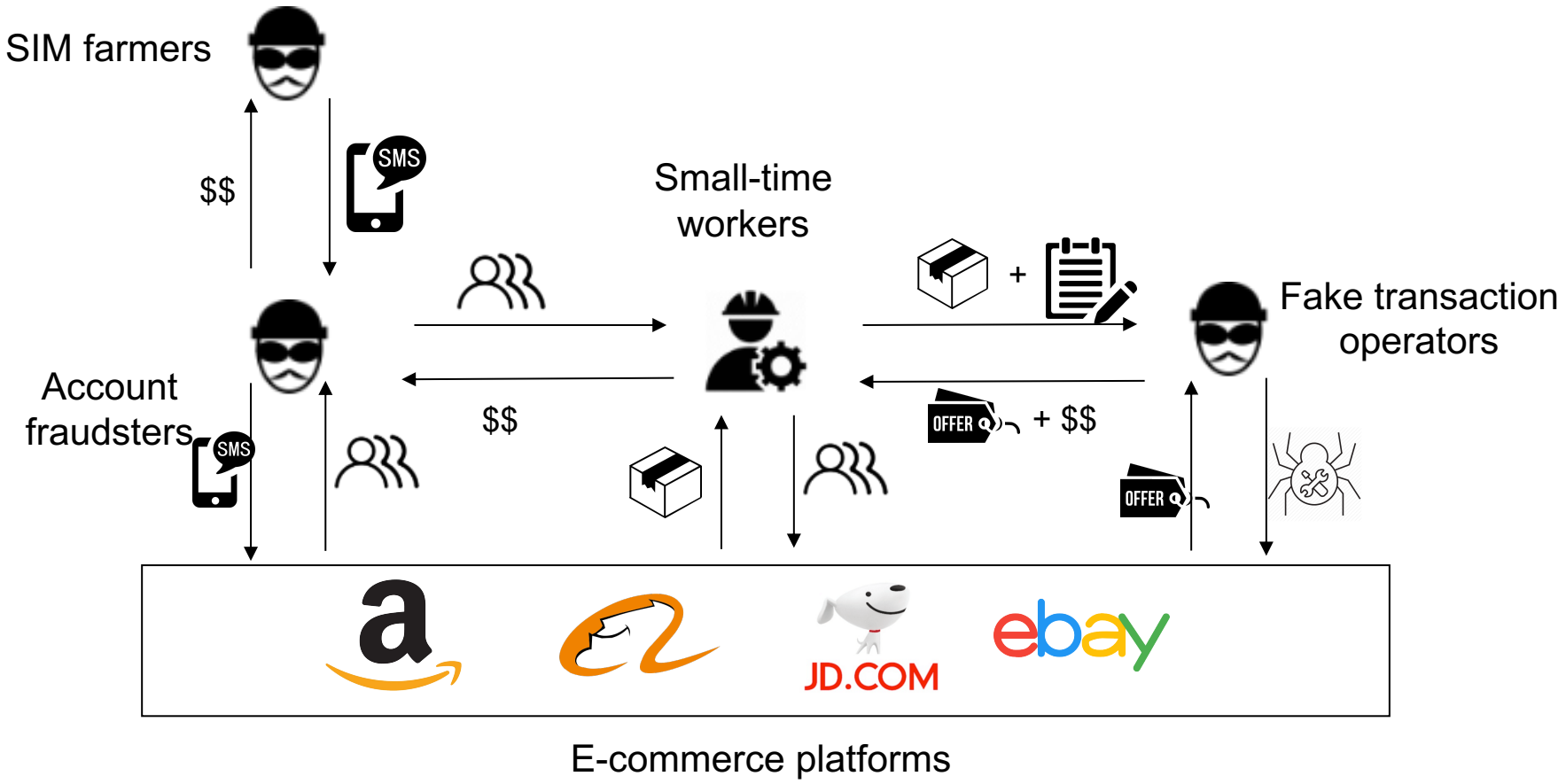


SIM farming

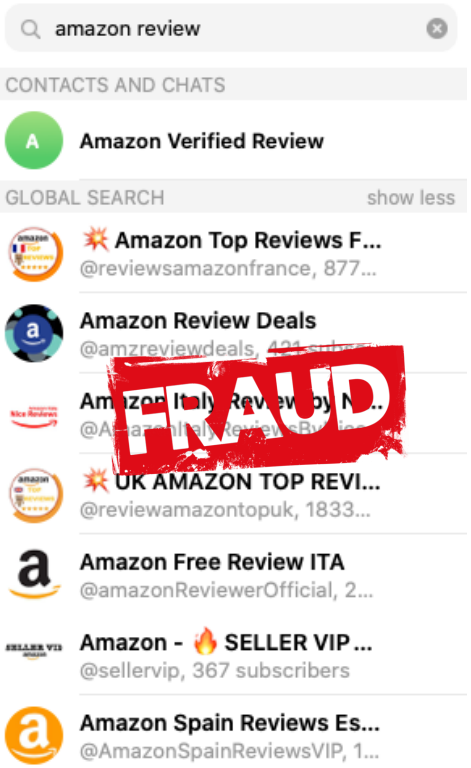
SIM Sources:
- VoIP cards
- …



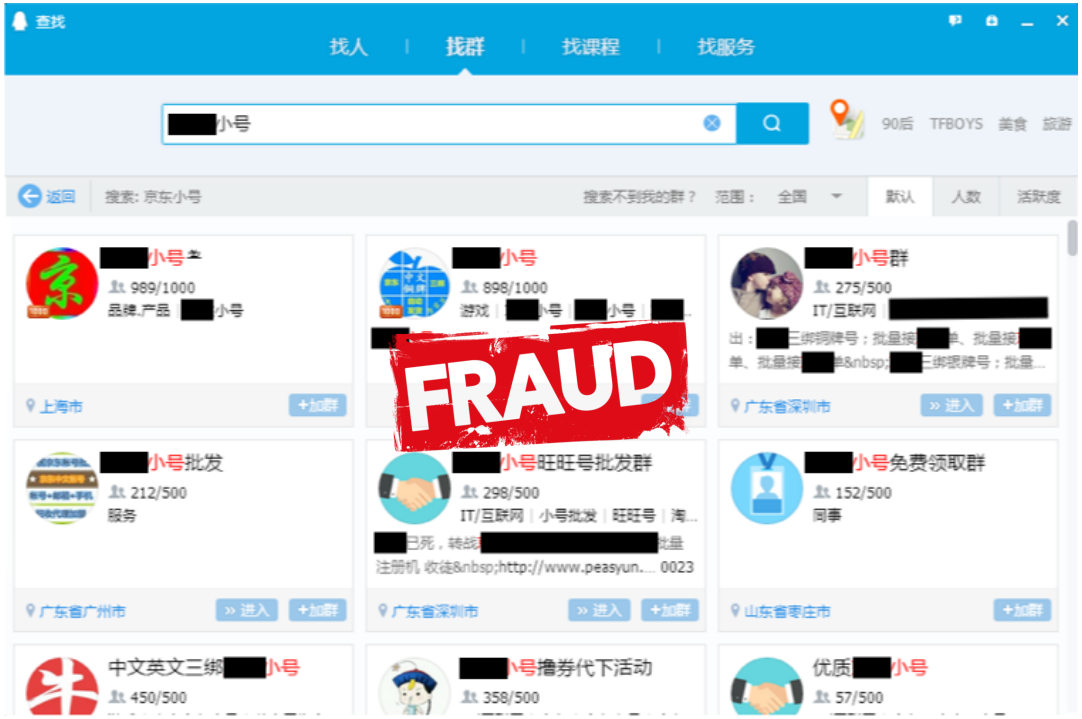
E-commerce fraud ecosystem



E-commerce fraud groups

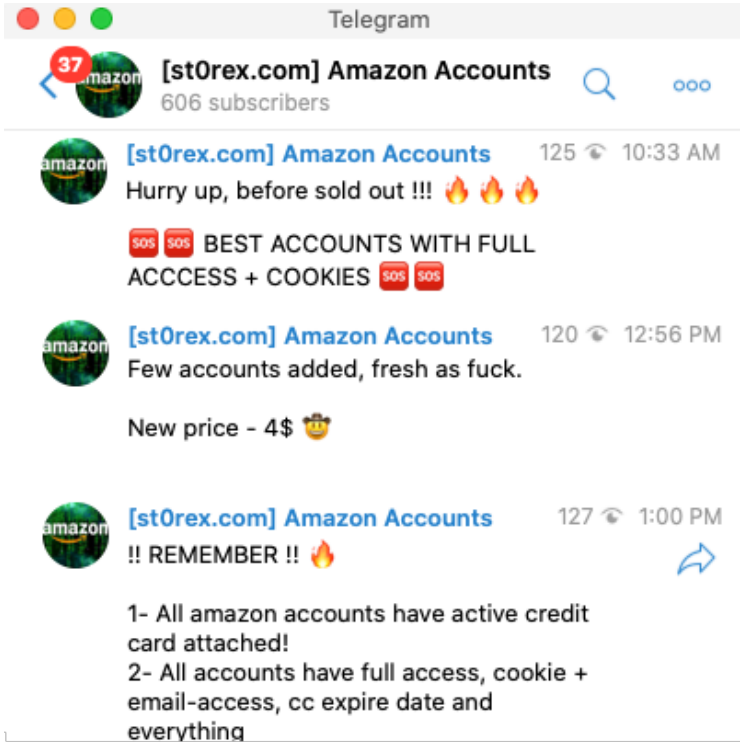


Fake review groups
on Telegram



Fraud account groups
on QQ

E-commerce fraud group chat



Group chat

Threat intelligence gathering:
collecting **evidence-based** threat information about an existing or emerging threat

SIM farmers:

1) SIM card source
2) Gateway link/tool
3) Account merchants
4) Hack tools

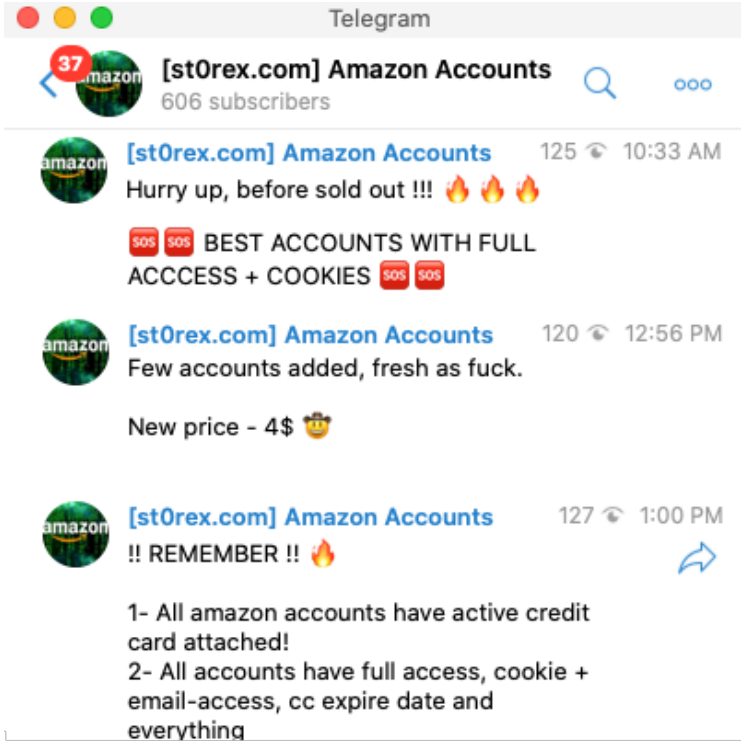
Fraud account merchants:

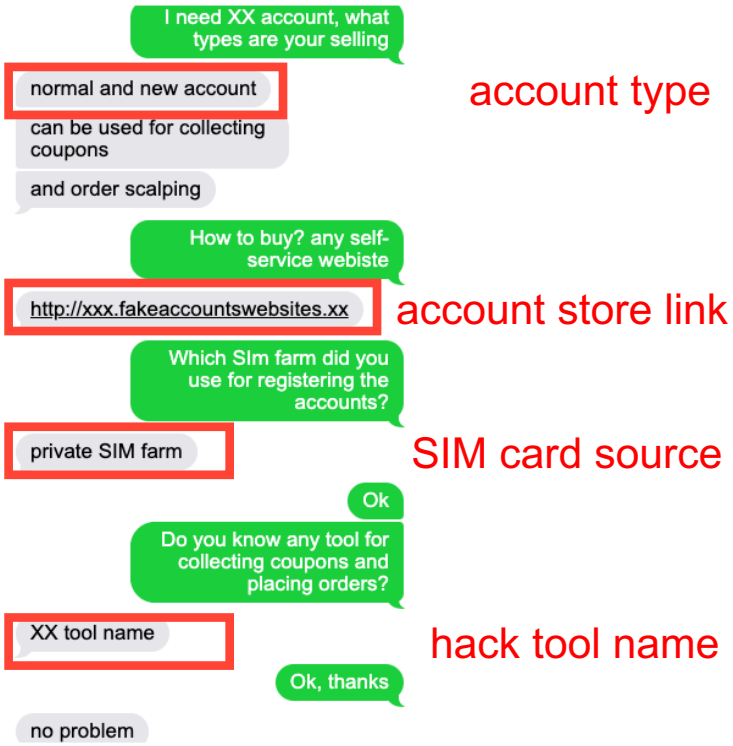
1) Account types
2) Store link
3) Payment method
4) SIM card source
5) Hack tools
6) Fraud order tasks

Fraud account operators:

1) Fraud order tasks
2) Shipping address
3) Report link
4) Hack tools
5) Account merchants

Group chat V.S. individual chat





Group chat V.S. Individual chat

Intelligence gathering challenges

- **Active** intelligence gathering
  - useful intelligence is only shared through one-on-one conversation
  - the number of new fraudsters keep growing

Intelligence gathering challenges

- **Active** intelligence gathering
  - useful intelligence is only shared through one-on-one conversation
  - the number of new fraudsters keep growing

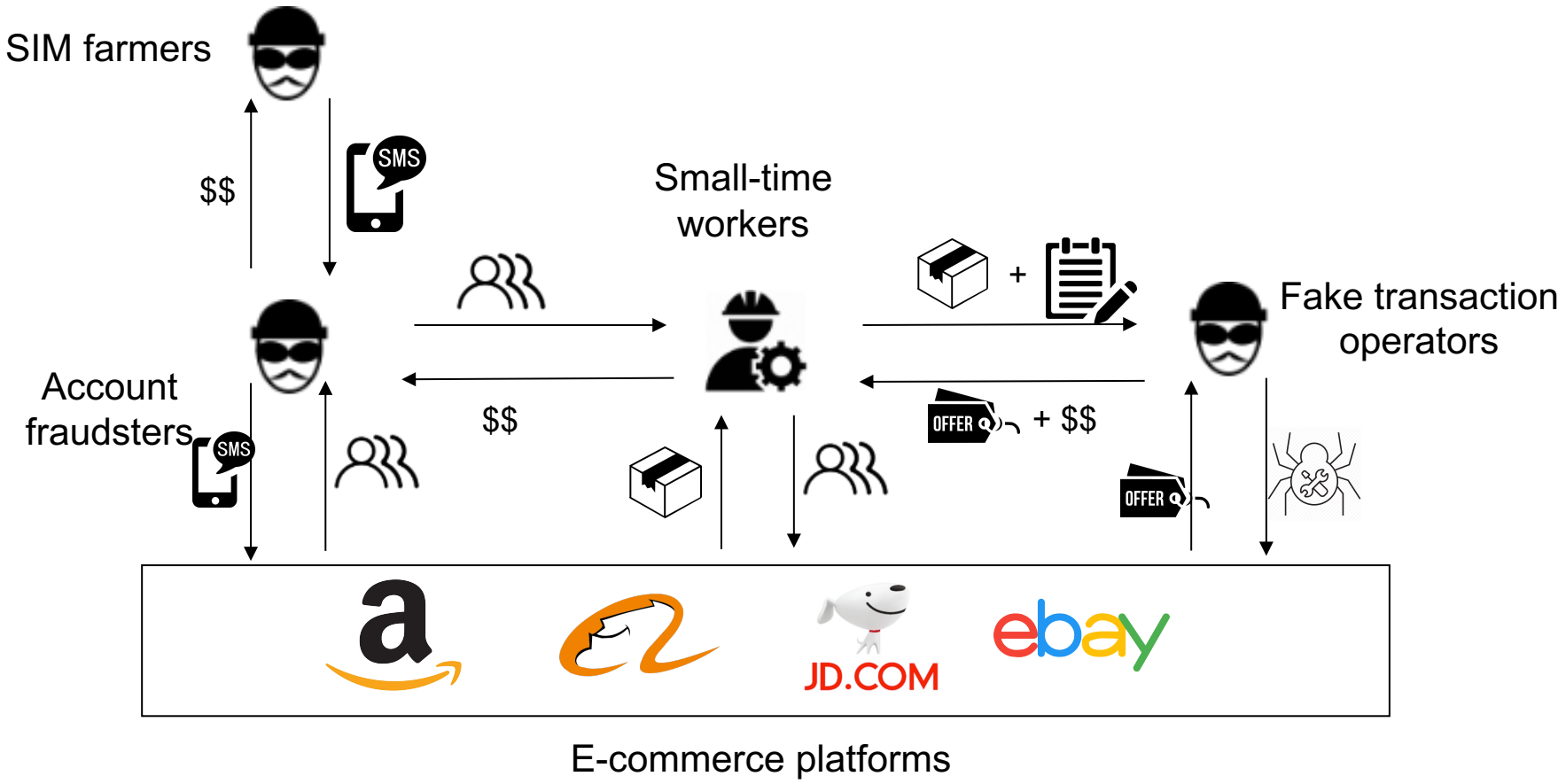
- Automated conversation with fraudsters
  - existing chatbots can not collect e-commerce threat intelligence
  - how to strategically lead the fraudsters to discuss the target threat intelligence is complicated

Aubrey

Autonomous chatbot for intelligence discovery

- **first** autonomous conversation system for **active** threat intel. gathering from e-commerce miscreants
- **effectively** extract great number of valuable **fraud-related artifacts**
- **new insights** into the e-commerce fraud ecosystem

Information exchange



Observation

Hi there

I need XX account, what types are you selling

normal and new account

can be used for collecting coupons

and order scalping

How to buy? any self-service website

http://xxx.fakeaccountswebsites.xx

E-commerce fraudster

Which Sim farm did you use for registering the accounts?

private SIM farm

Ok

Do you know any tool for collecting coupons and placing orders?

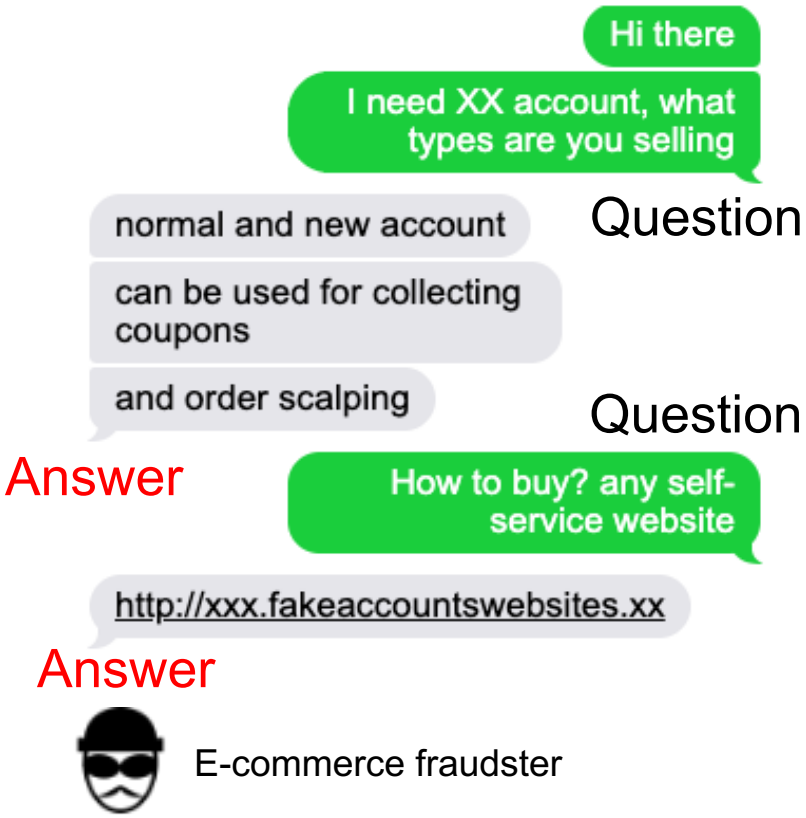
XX tool name

Ok, thanks

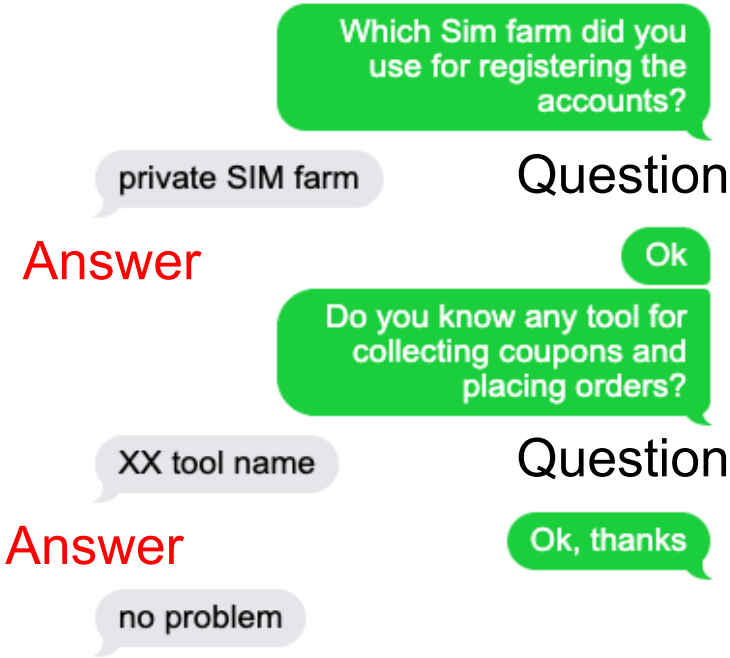
no problem

Small-time worker

Observation

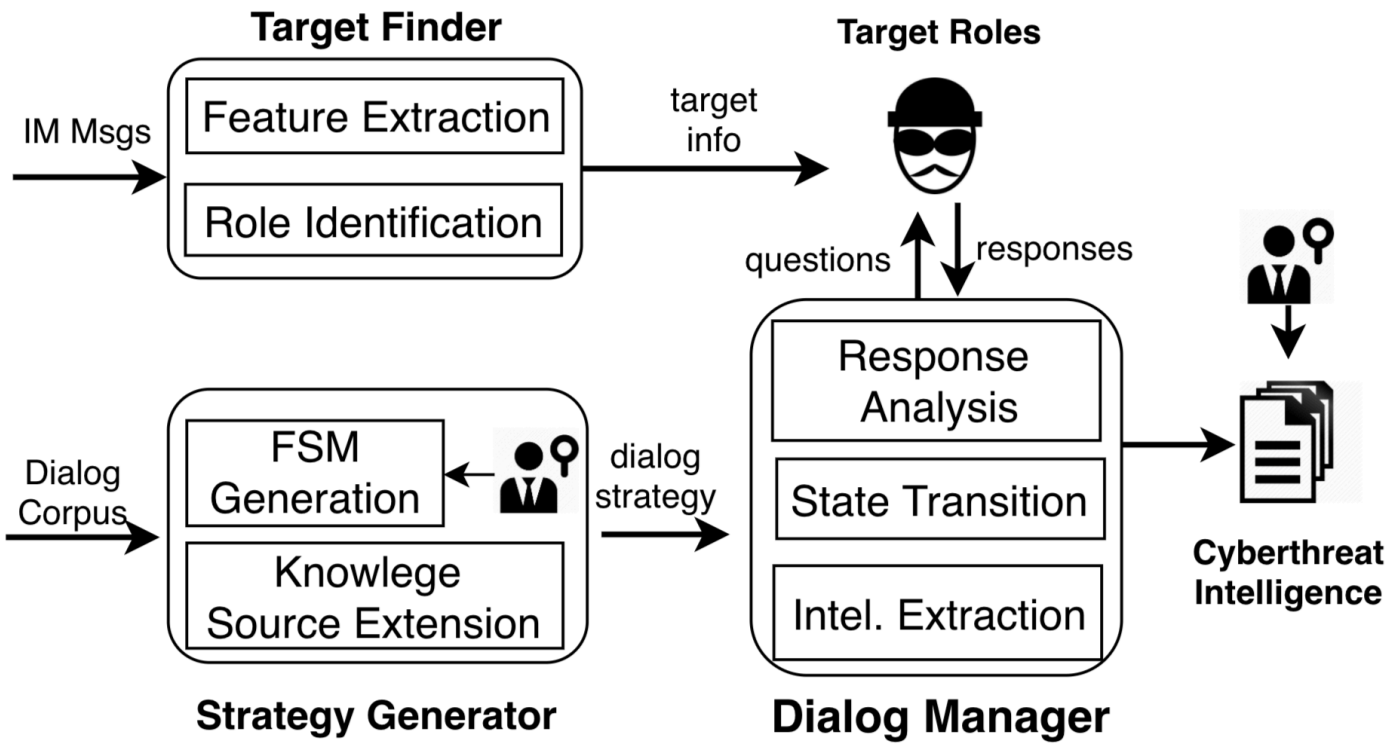


E-commerce fraudster

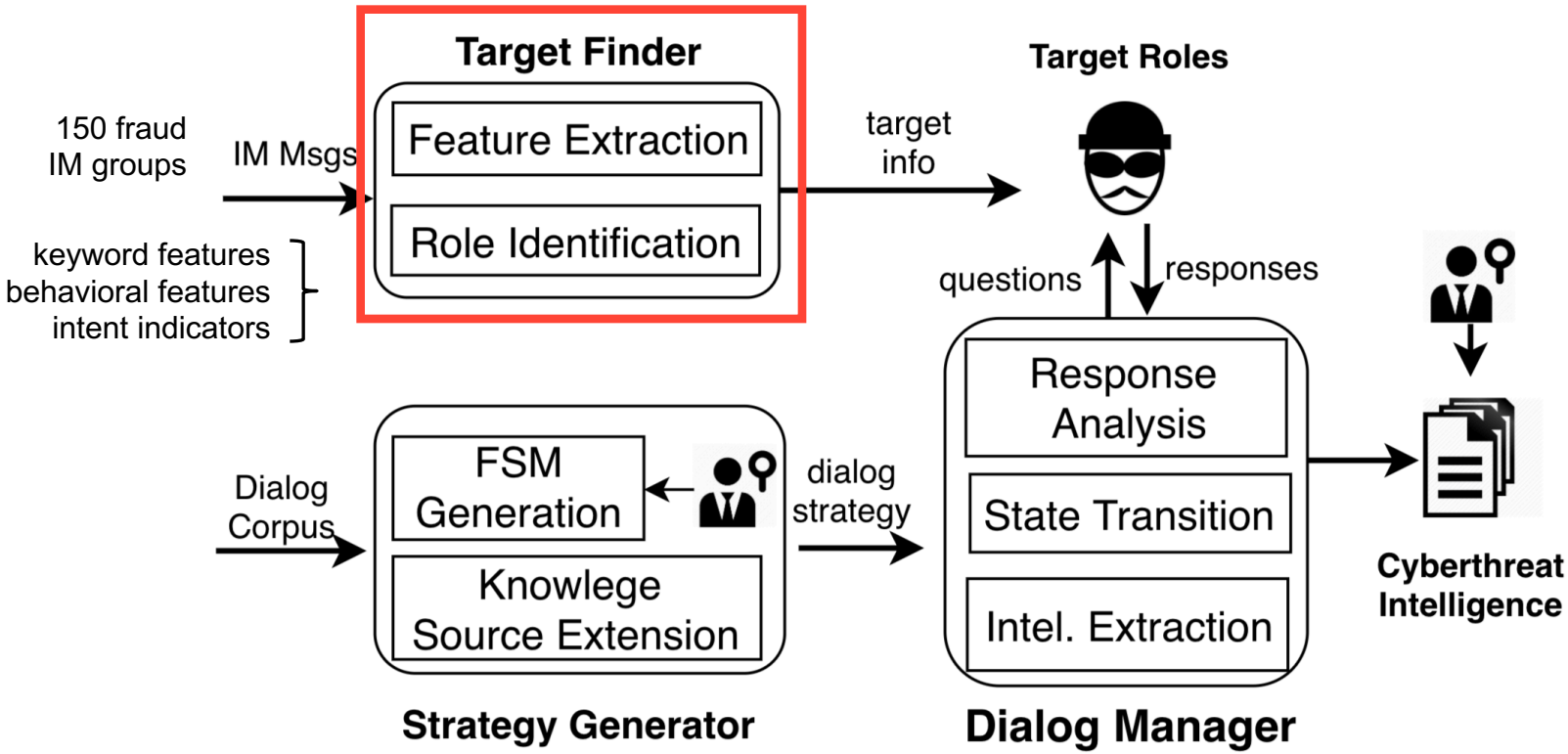


Small-time worker

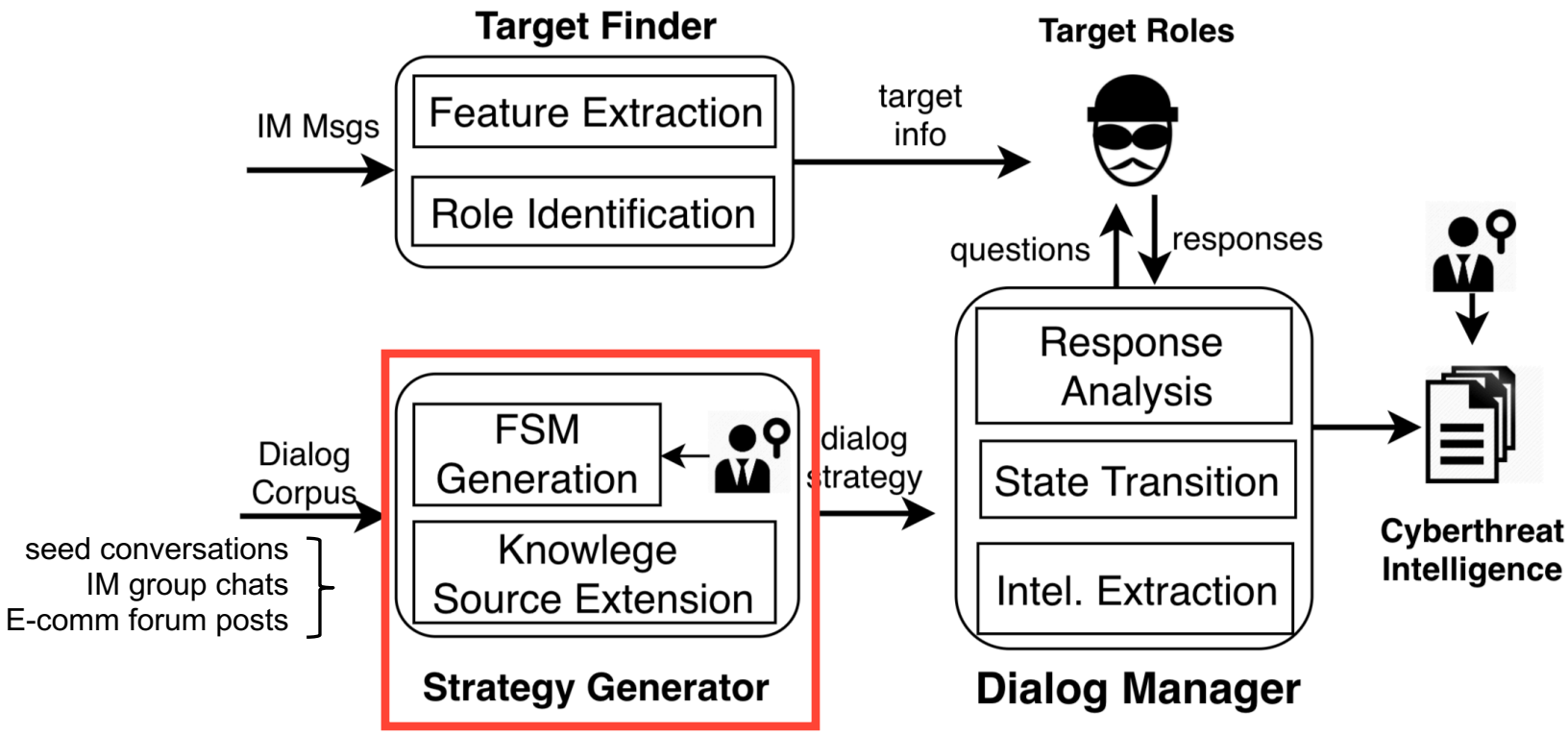
Architecture



Target Finder



Strategy Generator



FSM definition

5-tuple: (S, R, δ, s_0, E)

S: set of states, question Aubrey can send to the target roles

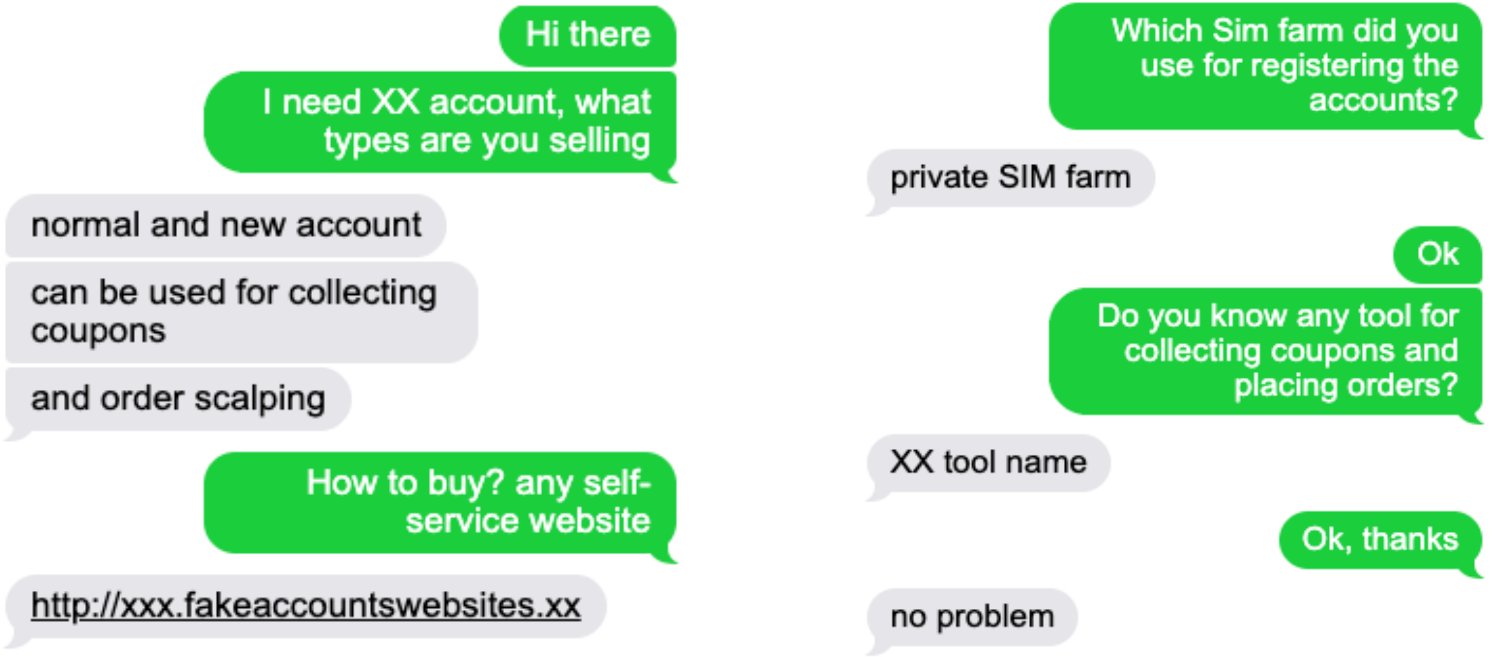
R: set of responses from the target roles

δ: $S \times R \rightarrow S$, state transition function, decide the next state

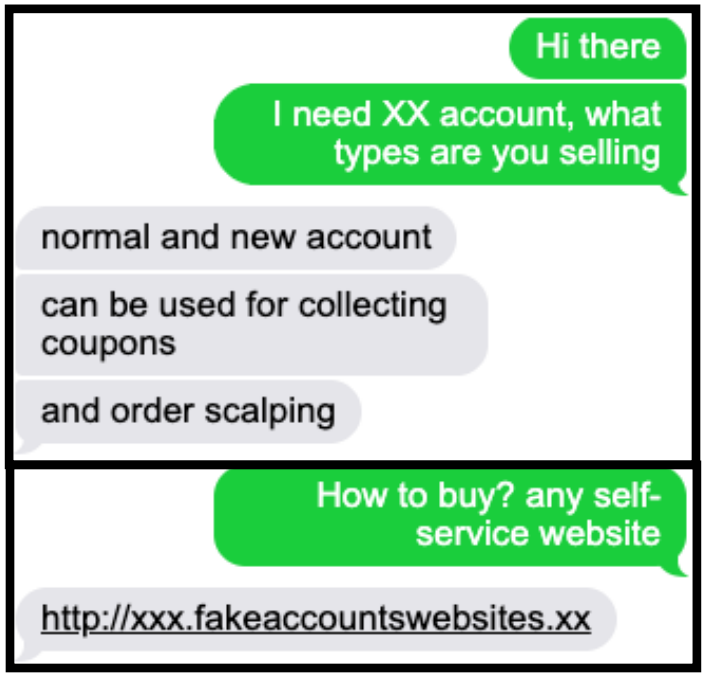
s_0: start state

E: end state

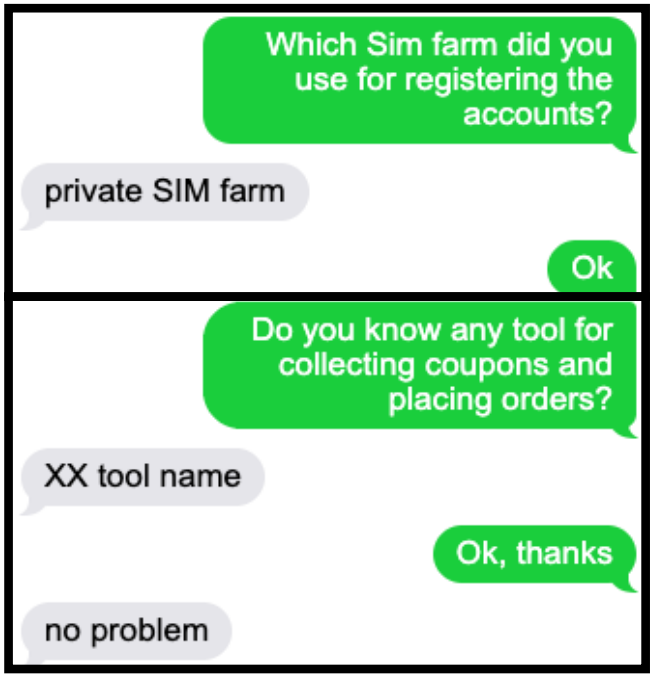
Seed conversation



Segmentation



dialog blocks

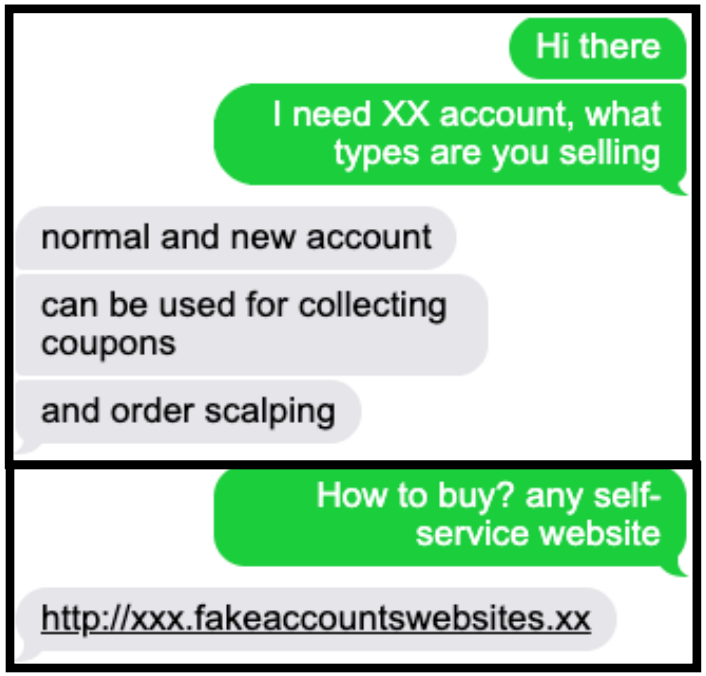


Seed conversation + text clustering

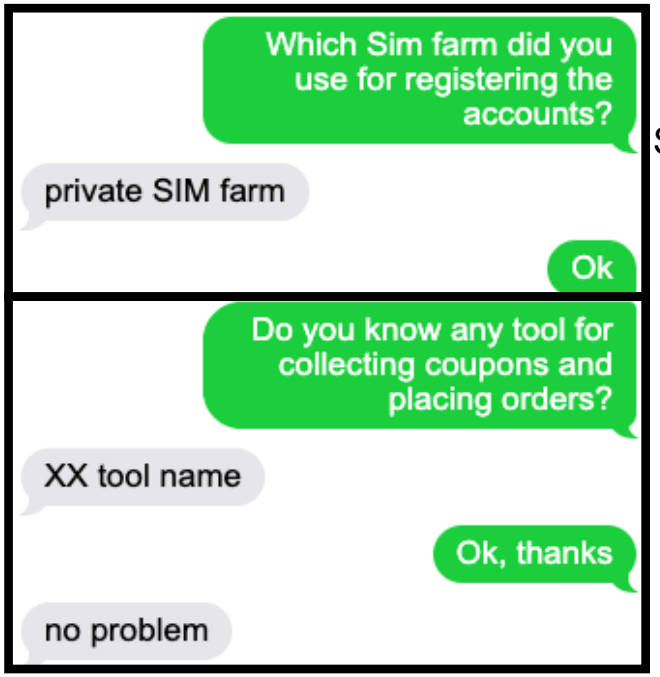
Topic detection

account types

storelink



dialog blocks

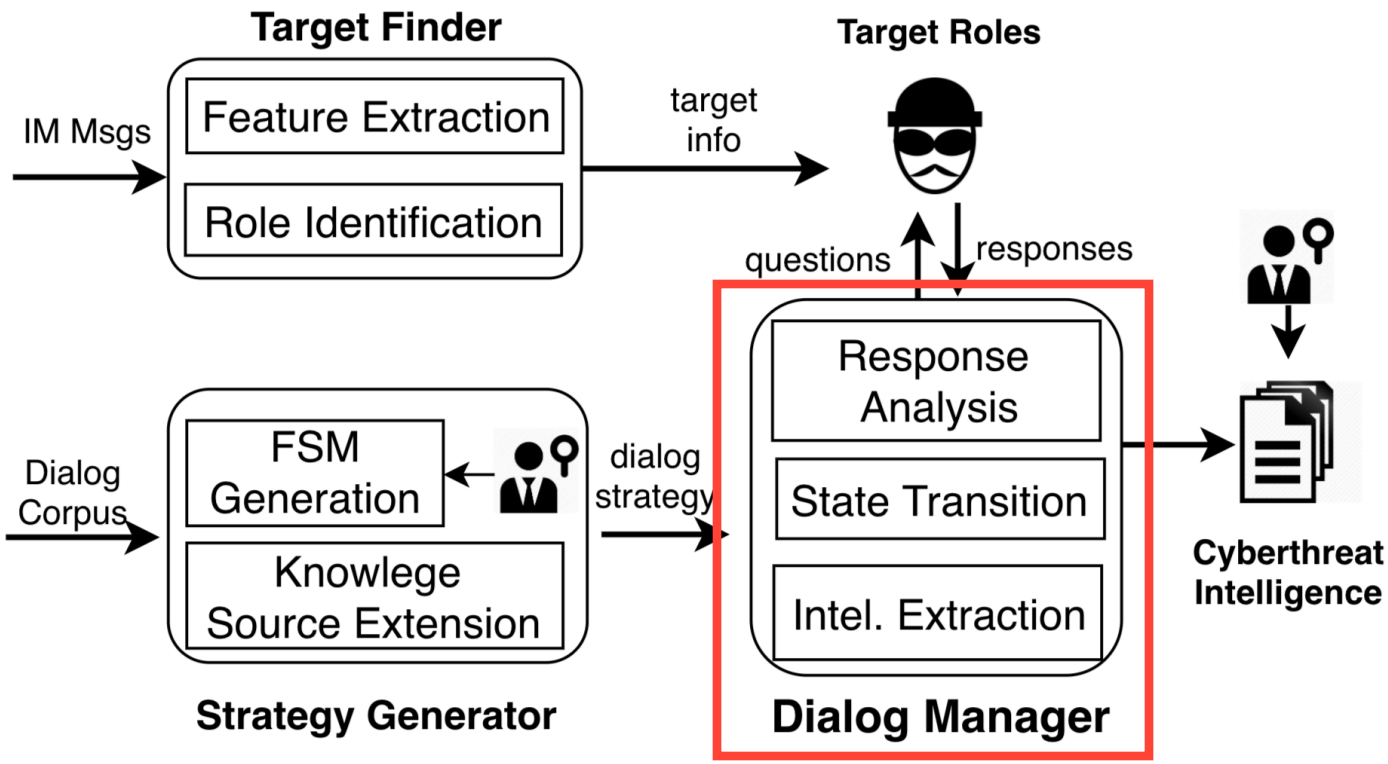


SimSource

Cross-role

topic identification + Seed conversation + text clustering

Dialog Manager



Retrieval model

- FSM for retrieval model

Current state × Response is interrogative

→ Retrieval model state

Evaluation

470 miscreants 7,250 communication messages

Category	# miscreants	# interactions
SIM farmers	185	2,900
Account merchants	130	2,350
Fraud order operators	155	2,000

Threat intelligence analysis

Category	Obtained intelligence	Extended intelligence
SIM farmers	40 SIM gateways, 36 payment intel. 16 SIM card sources and inventory intel. 8 fraud account websites 1 bonus hunting automated tool	323K fake phone numbers - 15 fraud account types, 8 payment intel. -
Account merchants	38 account trading websites 25 types of fraud accounts 26 payment intelligence, 10 SIM gateways 5 bonus hunting automated tools	150 fraud accounts, 6 hosting platforms - 14K fake phone numbers 10 private APIs
Fraud order operators	65 targeted items 184 fraud order addresses 71 fraud order report links 4 fraud account websites 6 bonus hunting automated tools (same as above)	65 fraud order affiliates 8 fraud address patterns 5 hosting platforms 8 fraud account types, 4 payment intel. 10 private APIs (same as above)

E-commerce miscreants and corresponding threat intelligence

Intelligence from SIM farmers

Obtained intelligence	Extended intelligence
40 SIM gateways, 36 payment intel.	323K fake phone numbers
16 SIM card sources and inventory intel.	-
8 fraud account websites	15 fraud account types, 8 payment intel.
1 bonus hunting automated tool	-

90% were used for account registration
72% accounts were used to order online

Intelligence from Account merchants

Obtained intelligence	Extended intelligence
38 account trading websites	150 fraud accounts, 6 hosting platforms
25 types of fraud accounts	-
26 payment intelligence, 10 SIM gateways	14K fake phone numbers
5 bonus hunting automated tools	10 private APIs

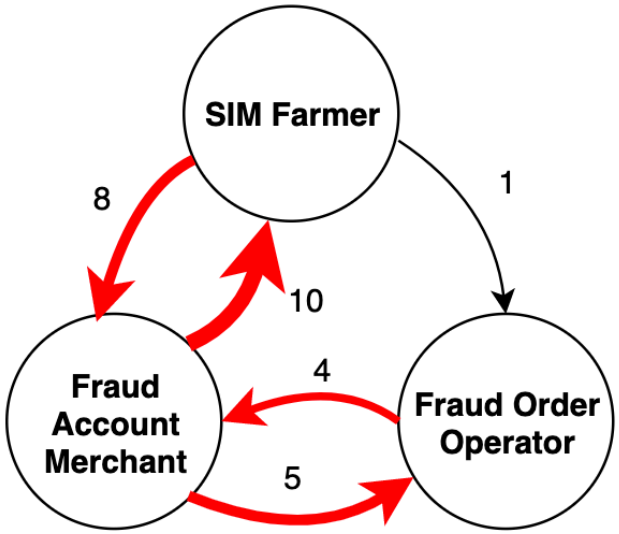
Abused private APIs and hack tools never been known before

Intelligence from Fraud operators

Obtained intelligence	Extended intelligence
65 targeted items	65 fraud order affiliates
184 fraud order addresses	8 fraud address patterns
71 fraud order report links	5 hosting platforms
4 fraud account websites	8 fraud account types, 4 payment intel.
6 bonus hunting automated tools (same as above)	10 private APIs (same as above)

Patterns
district + random name
district + random fruit name
district + random street + random letters
district + random street + specific letters
district + specific Chinese characters
district + specific last name + random first name
district + random street + specific Chinese characters
district + random street + specific last name + random first name

Hidden criminal infrastructures



Complicity of roles

Conclusion

Lesson learnt

- Chatbot is effective to study the cybercrime which are highly rely on crowdsourcing
- Account trading lies at the center of the fraud ecosystem, more effort should be put to mitigate the fraud account threats

Future work

- The current implementation of Aubrey is simple while effective;
- more complicated conversation (jargon identification), larger open domain corpora, hybrid model with human analyst involvement

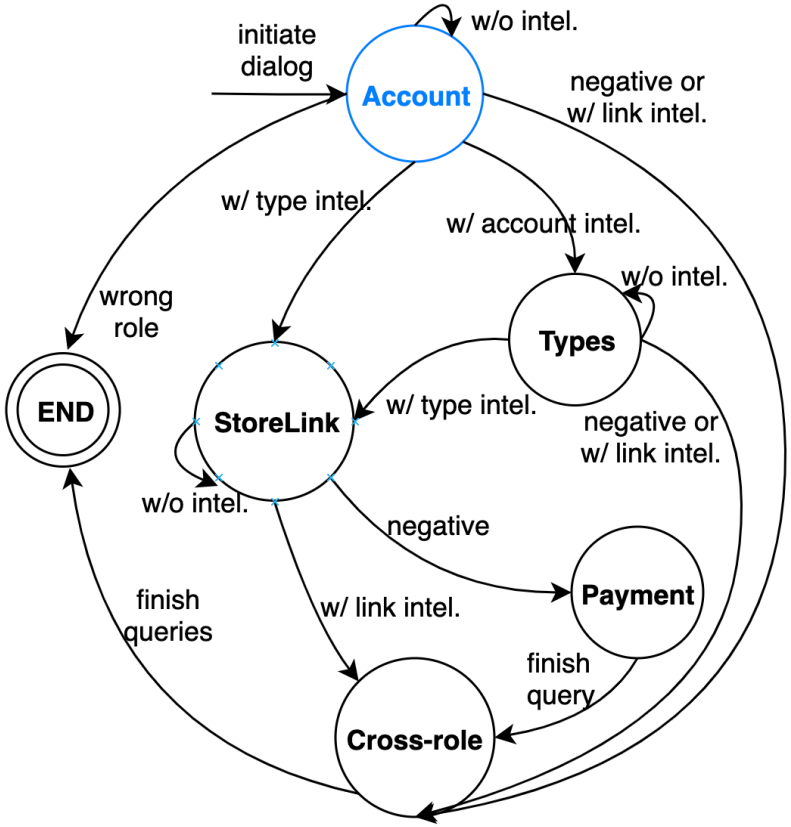
https://sites.google.com/view/aubreychatbot

Thank you !

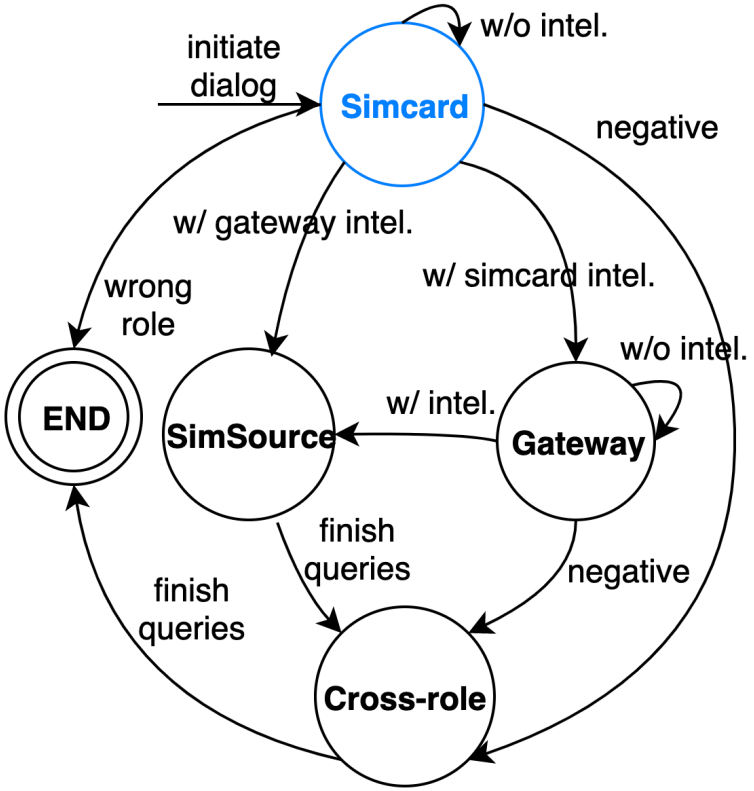
Discussion

- Scope
  - collected threat intel. is related to Chinese e-commerce platforms
- Generalization
  - with target intel. and domain-specific corpora, Aubrey can be re-trained to chat with other roles (drug dealers etc.) and languages
- Impact
  - fraud-related artifacts can be used as ground truth
  - fix exposed private APIs to raise the bar for automated abuse
  - stop fraudulent activities at the early stage

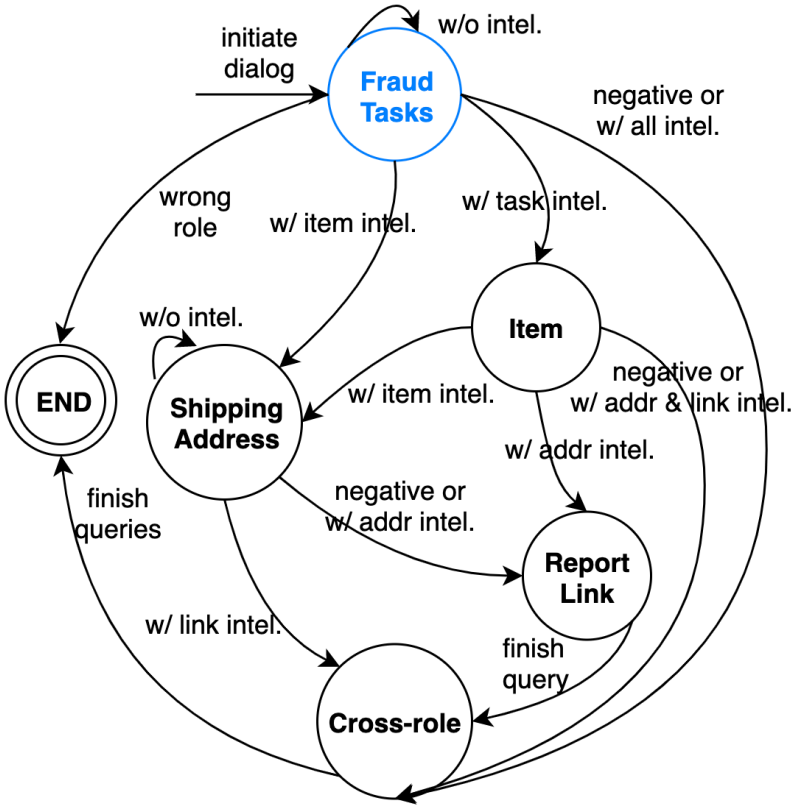
FSM for fake account trading



FSM for SIM farm and fake order operation

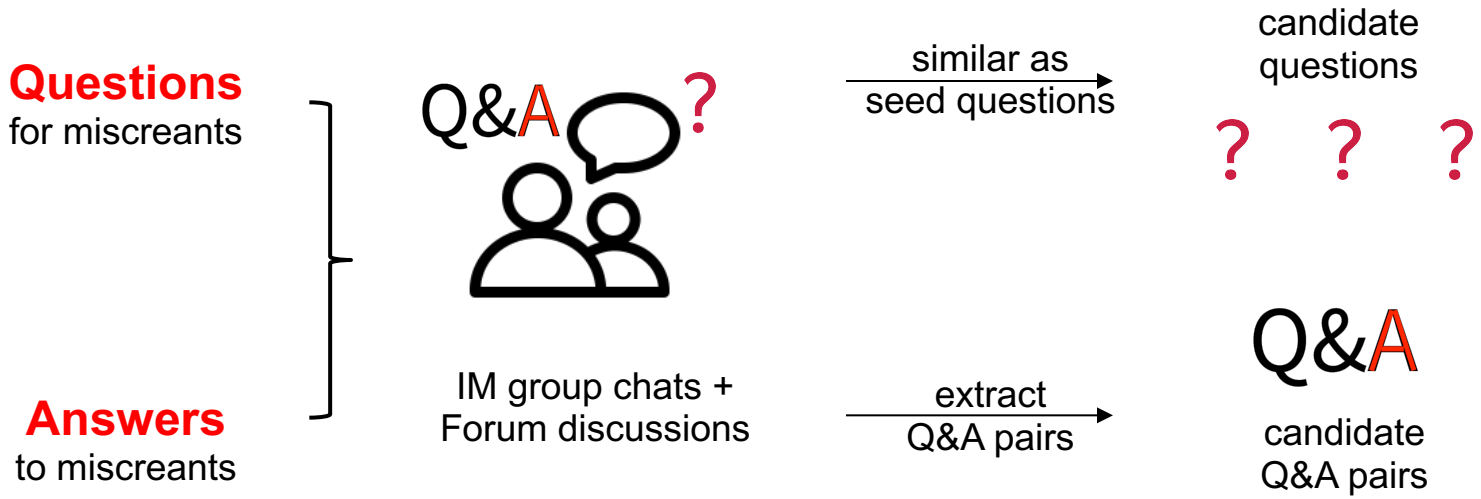


FSM for SIM farm



FSM for fake order operation

Knowledge source extension



Data collection

- Datasets

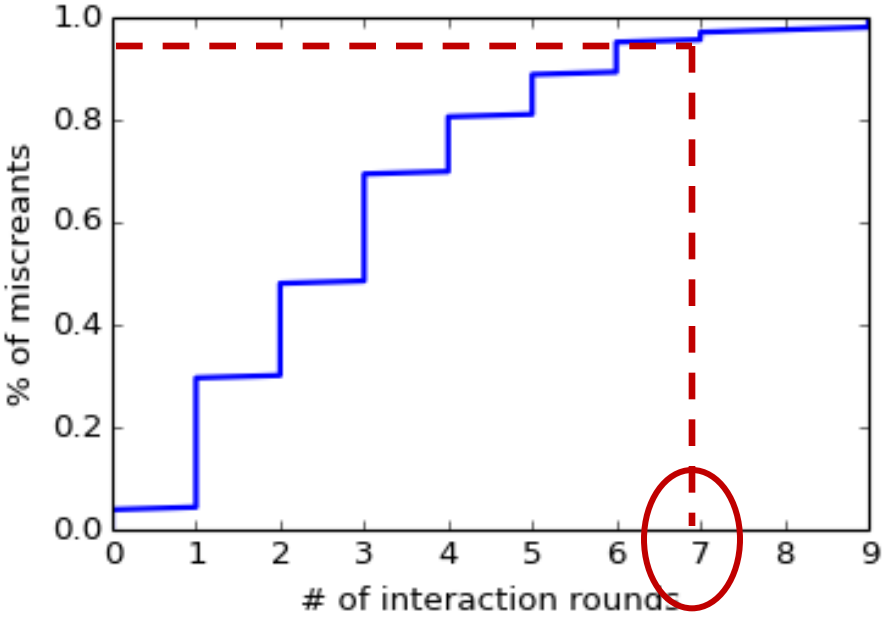
Dataset	# of raw data	# of dialog pairs
Seed conversation	800	200
IM group discussion	1 Million	50,000
Forum discussion	135,000	700,000

Evaluation

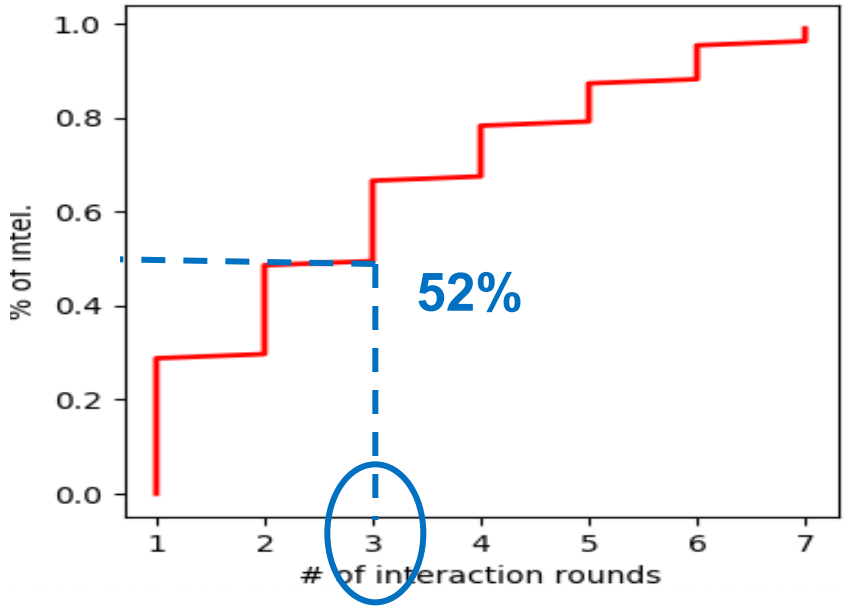
- Role identification classifier
  - Ground truth: 500 upstream, 180 downstream, 3,000 unrelated actors
  - Unknown set: 20,265 IM group members (from 150 IM groups)
  - Effectiveness:
    - upstream: 87.0% precision, 91.2% recall
    - downstream: 81.1% precision, 95.6% recall
    - upstream actor: 89.0% precision, 92.8% recall
    - overall: 86.2% F1 score

  1,044 SIM farmers, 700 account merchants, 2,648 fraud order ops
- Accuracy
  - 545 chat attempts, 470 responded (185 SIM farmers, 130 account merchants, 155 fraud order operators);
  - one questioned Aubrey
  - 97.4% (458) accuracy

Effectiveness

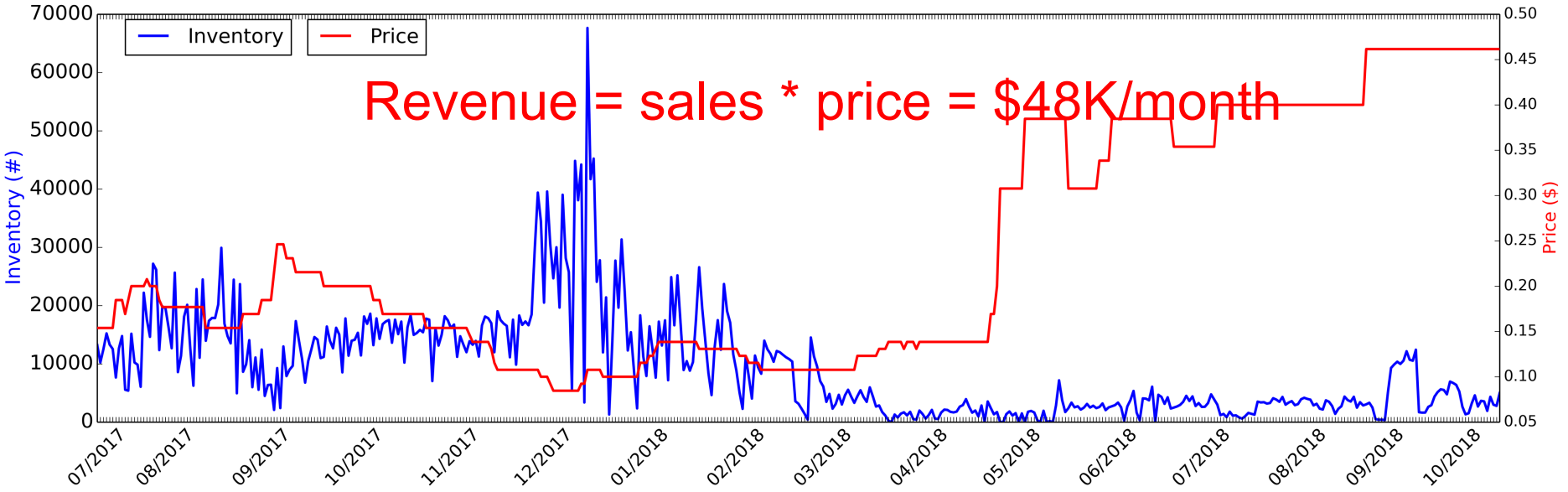


CDF of interaction round per miscreant



CDF of interaction round for intel. gathering

Case study



Account inventory and price tracking