On the State of OSN-based Sybil Defenses

David Koll*, Jun Li^, Joshua Stein^ and Xiaoming Fu*

***University of Göttingen, Germany*
^University of Oregon, Eugene, Oregon, United States

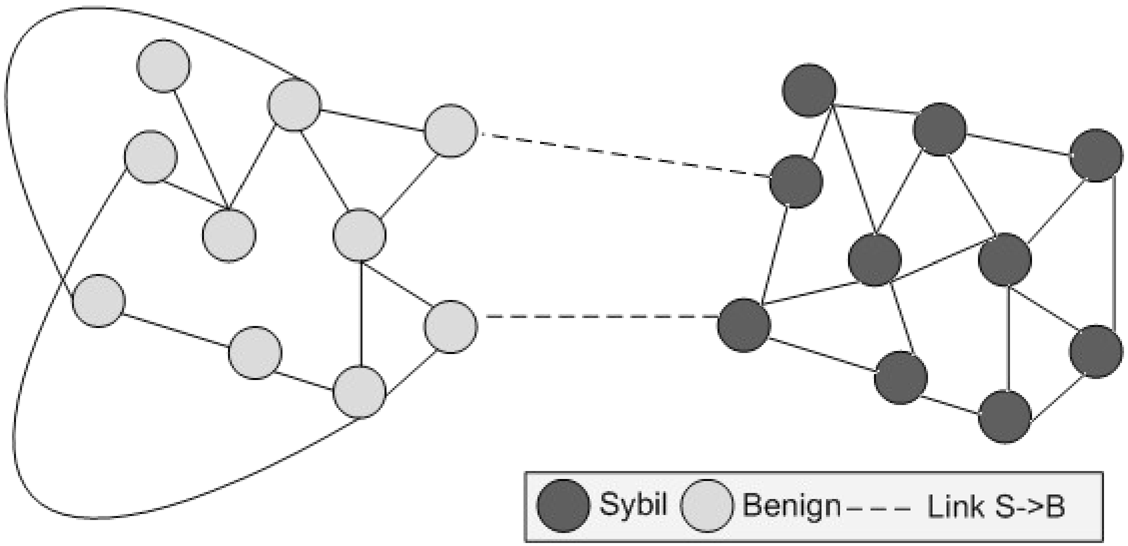
[koll|fu]@cs.uni-goettingen.de
[lijun|jgs]@cs.uoregon.edu

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

UNIVERSITY OF OREGON

Motivation

- **Sybil Attack**: injection of multiple forged identities into a target system with malicious intention
- **Current major research direction**: exploit Online Social Networks (OSNs) of users in target system
- **Idea**: it will be difficult for an attacker to create links to (become friends with) a benign user

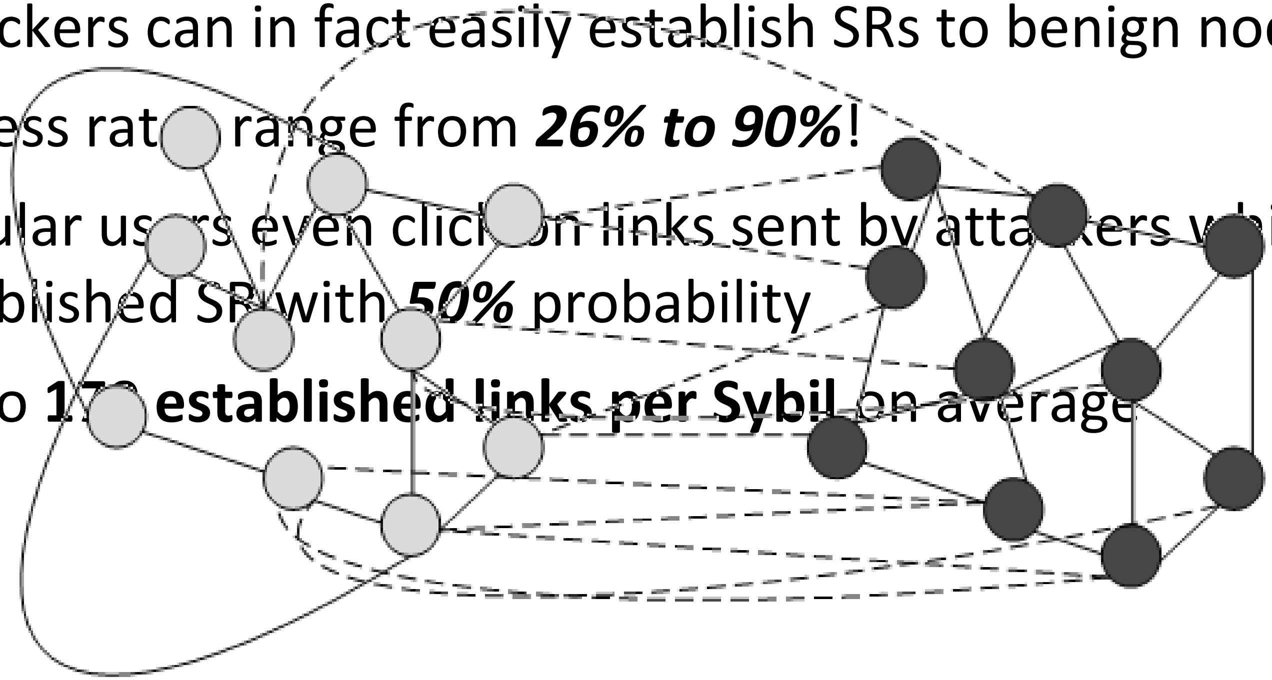


Motivation

Recent research suggests: ***Assumption invalid!***

1) Sybils can create only few links? [1,2,3]

- Attackers can in fact easily establish SRs to benign nodes, success rates range from ***26% to 90%***!
- Regular users even click on links sent by attackers which just established SR with ***50%*** probability
- Up to **17[illegible] established links per Sybil** on average

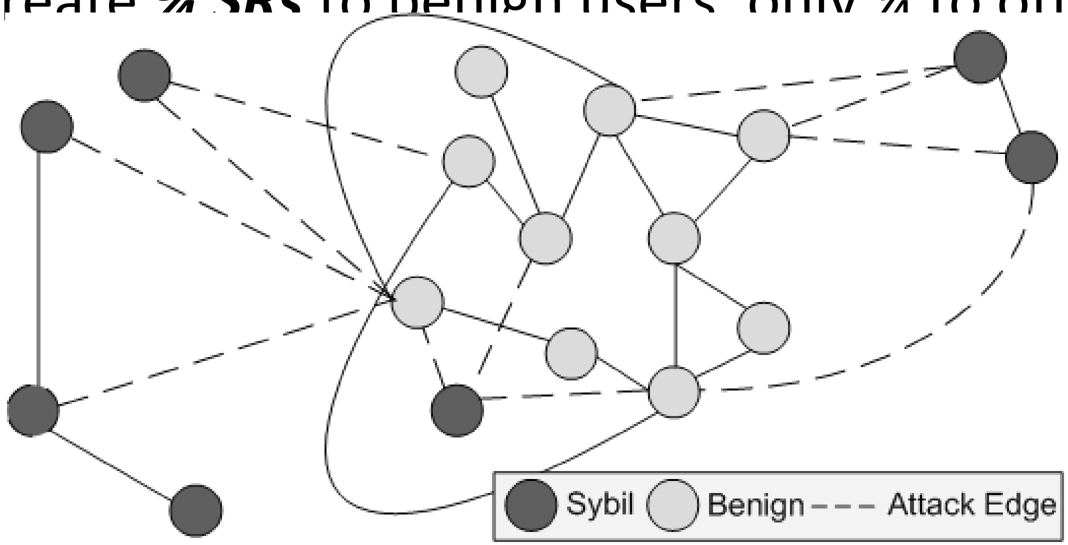


Motivation

Recent research suggests: ***Assumption invalid!***

2) Sybils keep among themselves? [2]

- Sybils create **¾ SRs** to benign users, only ¼ to other Sybils



GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

UNIVERSITY OF OREGON

Motivation

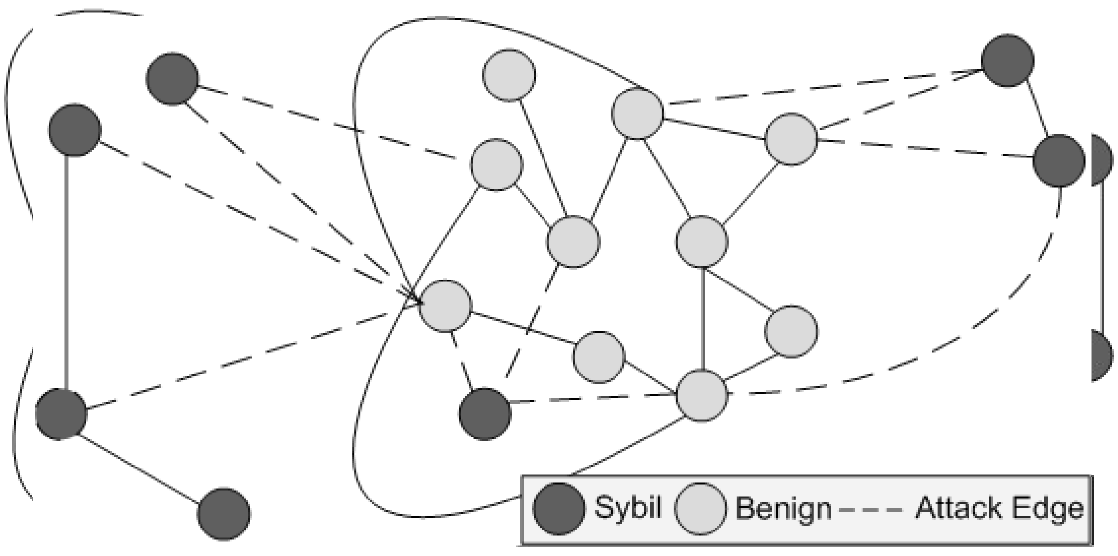
Recent research suggests: ***Assumption invalid!***

3) Attacker has to take initiative? [4,5]

- Simple attack strategies lure users into initiating contact with attacker.
- Socialbots can acquire ***hundreds*** of SRs to benign users ***per day, per profile***.
- Spammers on twitter gain hundreds of followers

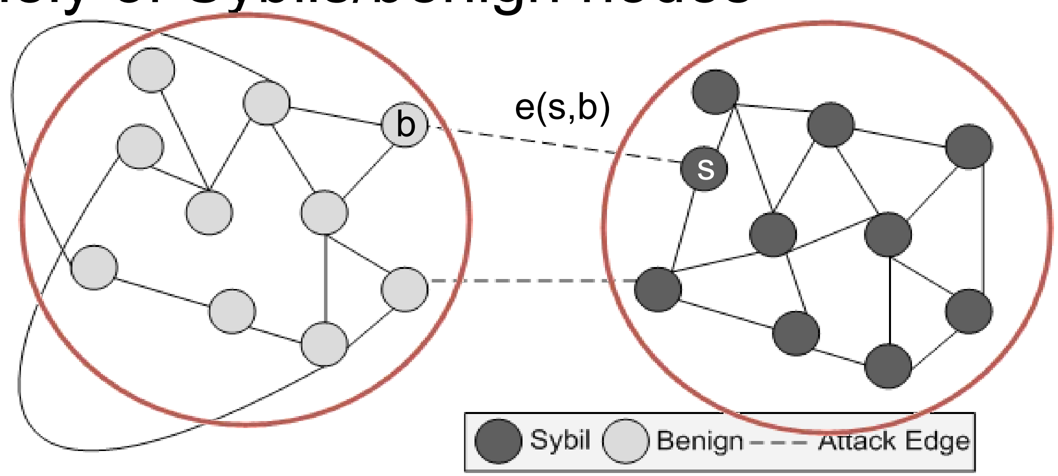
Motivation

- ***Our work:*** systematically analyze the State of the Art with regards to the new observations



Some Notations

- ***Sybil node:*** A forged identity controlled by the attacker
- ***Benign node/user:*** A regular, non malicious node/user
- ***Attack Edge***: An edge *e(s,b)* in the OSN graph *G=(V,E)* that connects a Sybil node *s* to a benign node *b*, i.e., a SR between *s* and *b*
- ***Sybil/Benign Community:*** A densely connected community consisting solely of Sybils/benign nodes



GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

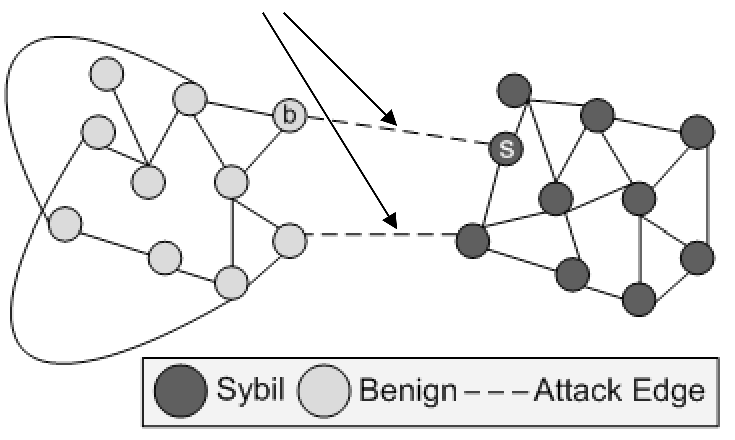
UNIVERSITY OF OREGON

OSN-based Sybil Defenses

- Two categories: *Sybil Detection (SD)* and *Sybil Tolerance (ST)* schemes
  - *SD*: Detect Sybils and exclude them from the system
    - e.g., SybilGuard/SybilLimit [NSDI'06/SP'08], SybilInfer [NDSS'09], SybilRank [NSDI'12], GateKeeper [INFOCOM'11]
  - *ST*: Accept that there are Sybils – tolerate them and mitigate their impact instead
    - e.g., Ostra [NSDI'08], SumUp [NSDI'09]

SD Approaches - Overview

- Most SD approaches use (modified) random walks to detect Sybils
  - Use bottleneck cut defined by the few attack edges



  - Random walk starting at *b* unlikely to cross to Sybil region, thus unlikely to end at/intersect with walk starting at *s*
  - Only exception: GateKeeper, uses ticket distribution

SD Approaches - Overview

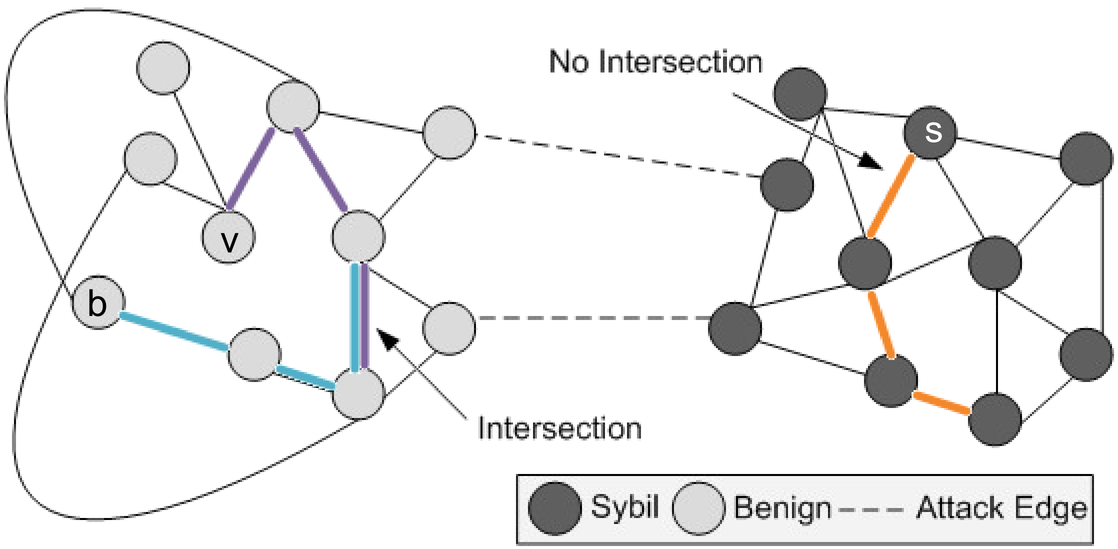
- Yes/no decision, whether suspect is admitted
- Basically the same idea over all approaches:

Low reachability of Sybils from honest users

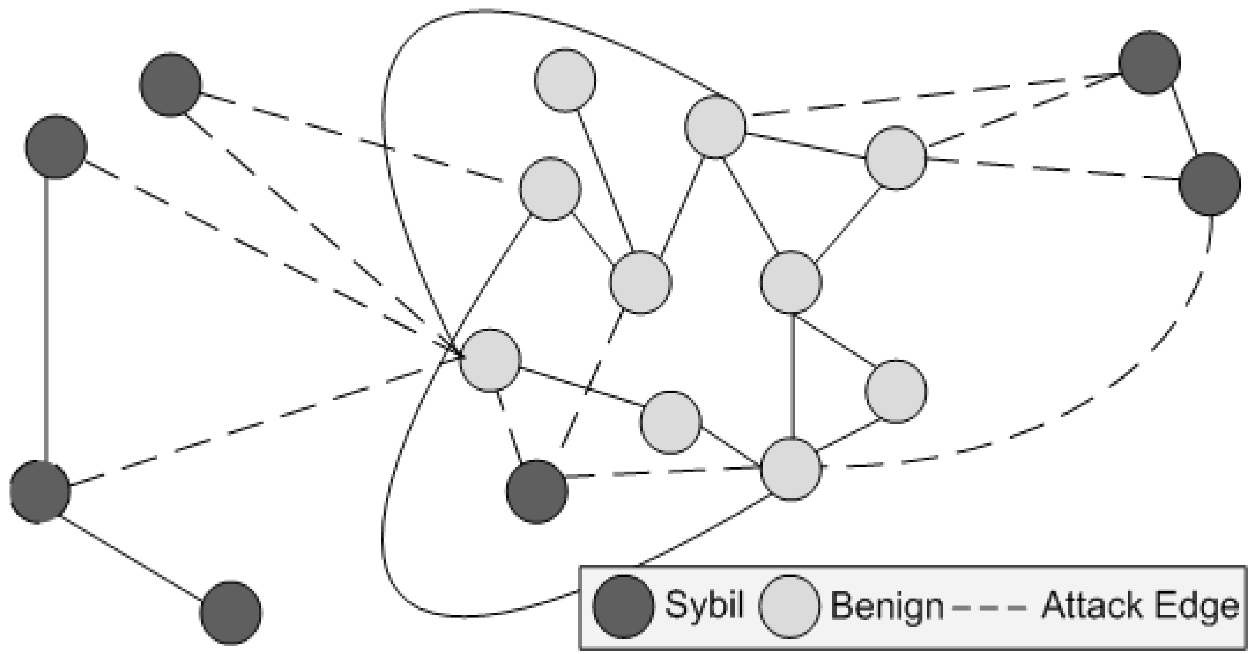
  - Random walks of Sybils should not intersect with honest users' walks (SybilGuard/Limit)
  - Sybils should have lower rank than honest nodes (SybilRank)
  - Sybils should obtain less tickets than honest nodes (GateKeeper)

SD Approaches – Example: SybilLimit

- Every node (suspect) has to be admitted by a verifier
- **Admission Concept:** Intersections of tails (last edge of the random walk)
  - **Idea:** Honest users will have a lot of intersections with honest verifiers...
  - ... while Sybils will not



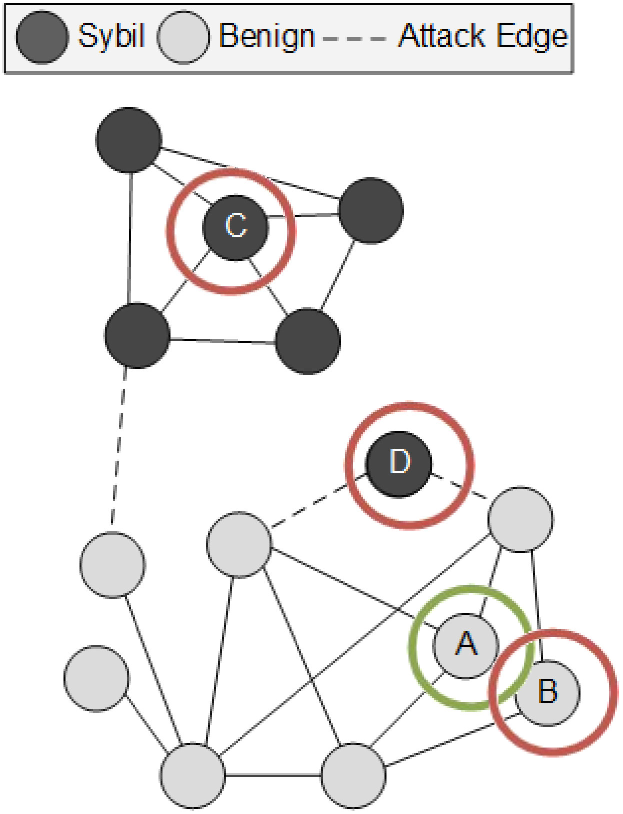
SD Approaches – Example: SybilLimit



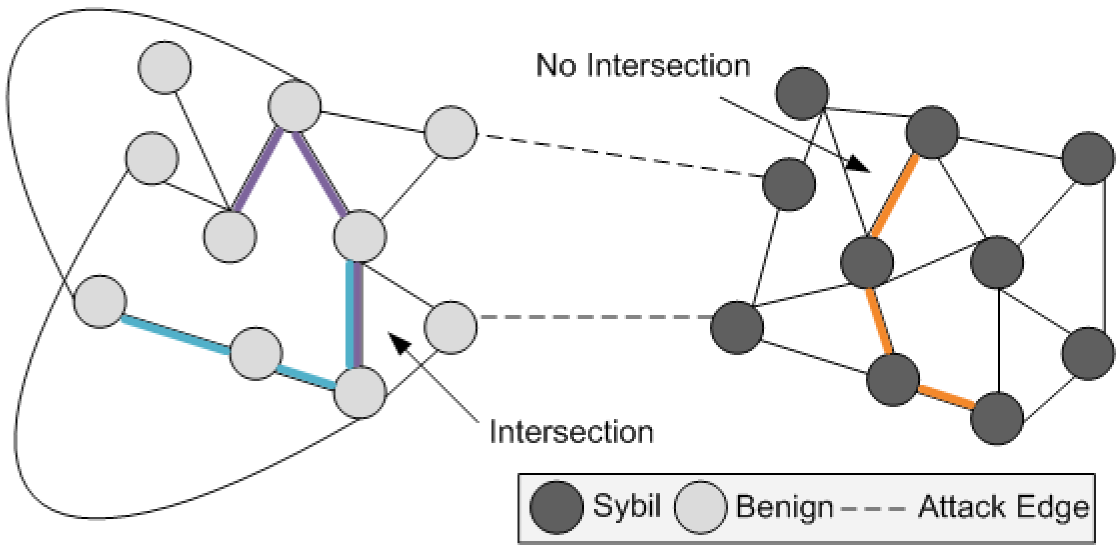
What now?

SD Approaches – Example: SybilLimit

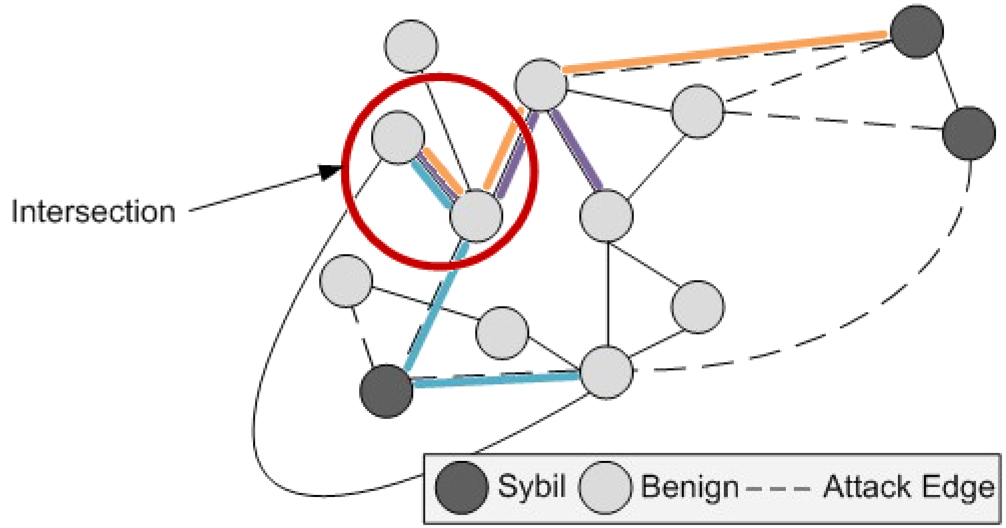
- Suspect gets admitted if there are intersections on the tails with a verifier
- ***Few attack edges:*** few intersecting tails between Sybils and honest nodes (e.g., walks starting at A are not likely to have intersecting tails with those at the Sybil C)
- ***More attack edges:*** SybilLimit can not distinguish between Sybils and honest nodes (e.g., nodes B and D)



SD Approaches – Example: SybilLimit



SD Approaches – Example: SybilLimit



SD Approaches - Overview

- We observe the same problem in every approach

Low distinguishing ability of the schemes

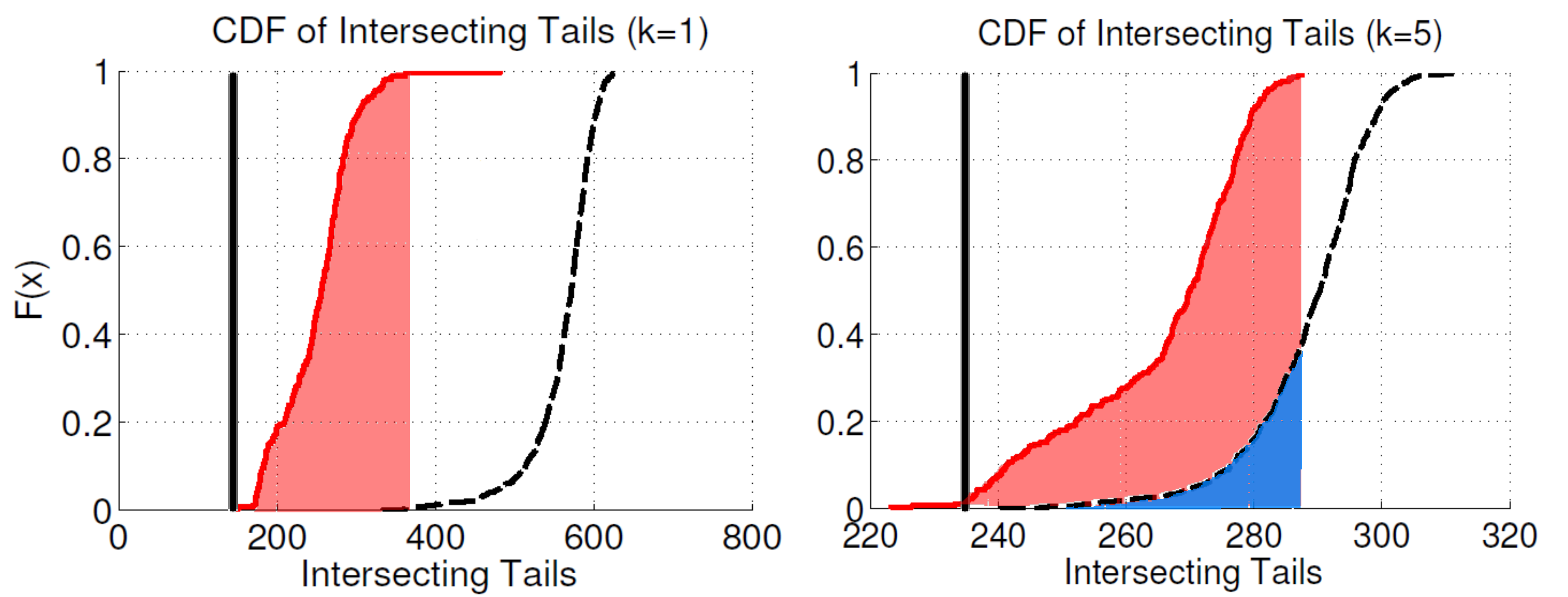
  - Significant difference in intersections... (SybilGuard/Limit)
  - ...or obtained rank... (SybilRank)
  - ...or ticket count (GateKeeper) no longer given

SD Evaluation - Methodology

- Datasets with different characteristics (*no dependency on dataset*) :
  - 1 synthetic, 1000 nodes, 2000 links, scale-free topology
  - 1 Facebook, 65000 nodes, over 3 million links
- Attackers are not allowed to deviate from System protocol
  - i.e., evaluate their gain by position in graph alone!
- Main parameter:
  - Number of attack edges per Sybil, *k*
  - Edge placement:
    - Random: each Sybil places *k* edges to benign nodes randomly
    - 100 different, independent placements to avoid biased results

SD Evaluation - SybilLimit

- Original SybilLimit: virtually admits every Sybil when k=1
  - Not surprising: guarantee of O(*log* n) admitted Sybils *per attack edge*
- Modification: try to distinguish on *number of intersecting tails*



SD Evaluation - Commonalities

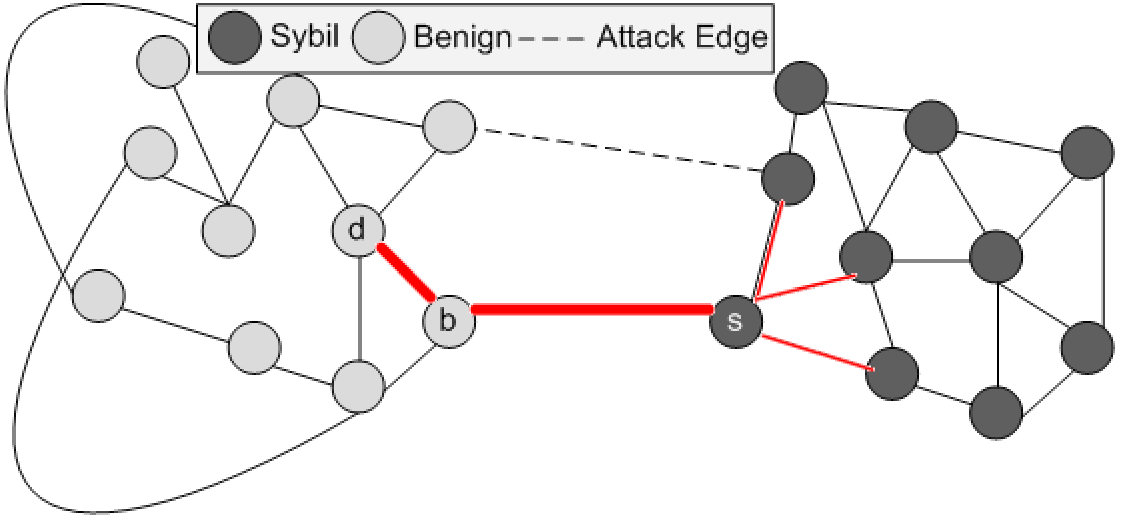
- Same problem in all defenses: Sybils are able to outperform large fractions of honest nodes with little effort
  - SybilInfer, SybilRank, GateKeeper: 1-2 attack edges sufficient
  - Effort can even be reduced by more intelligent placement strategies
  - Confirms the *low distinguishing ability*

ST Approaches - Overview

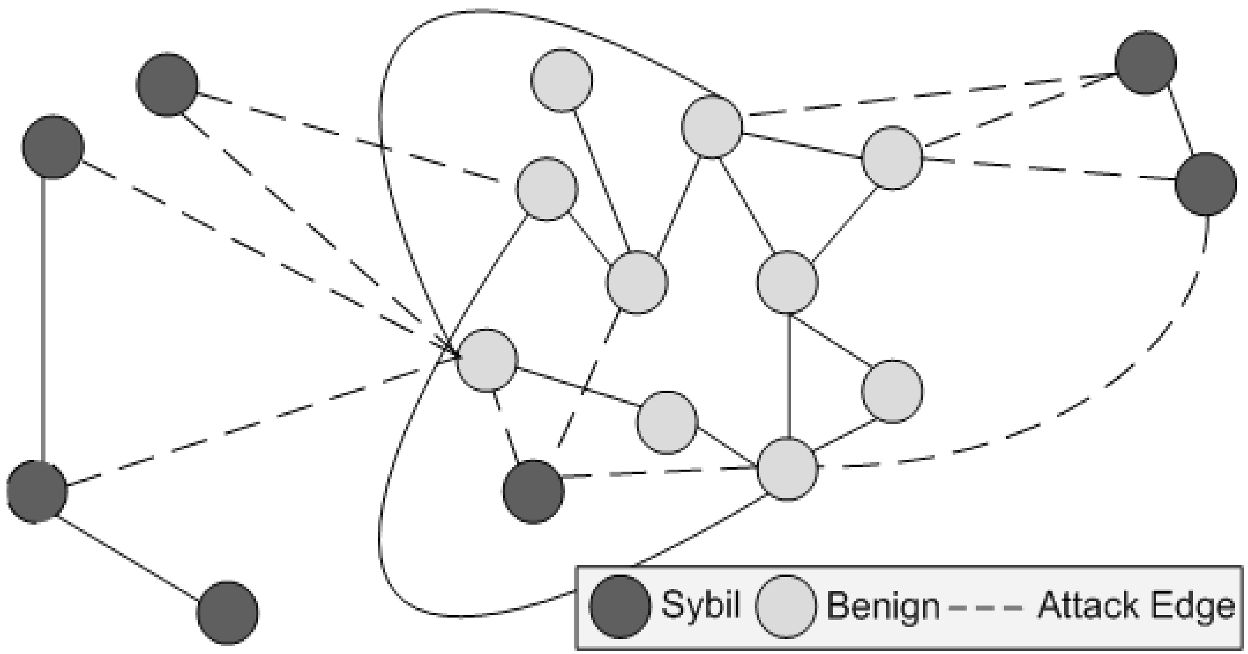
- ST approaches try to ***limit the impact*** of each admitted Sybil
- Most approaches are built on credit networks
  - A message can only be sent along a path if every link on the path has credit available
  - ST approaches exploit that credit should deplete quickly on attack edges

ST Approaches – Example: Ostra

- Assigns credits to links; messages may only be routed over links with credit
- If message is labeled as unwanted, credit on the path is deducted
- Sybils have to use few attack edges to transmit their spam



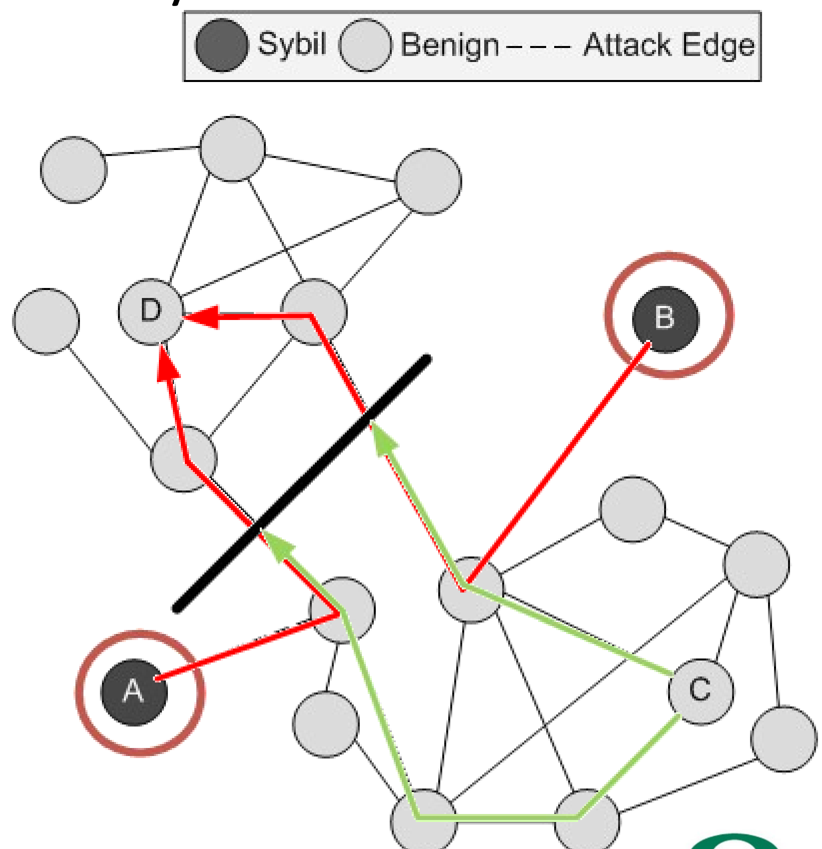
ST Approaches – Example: Ostra



What now?

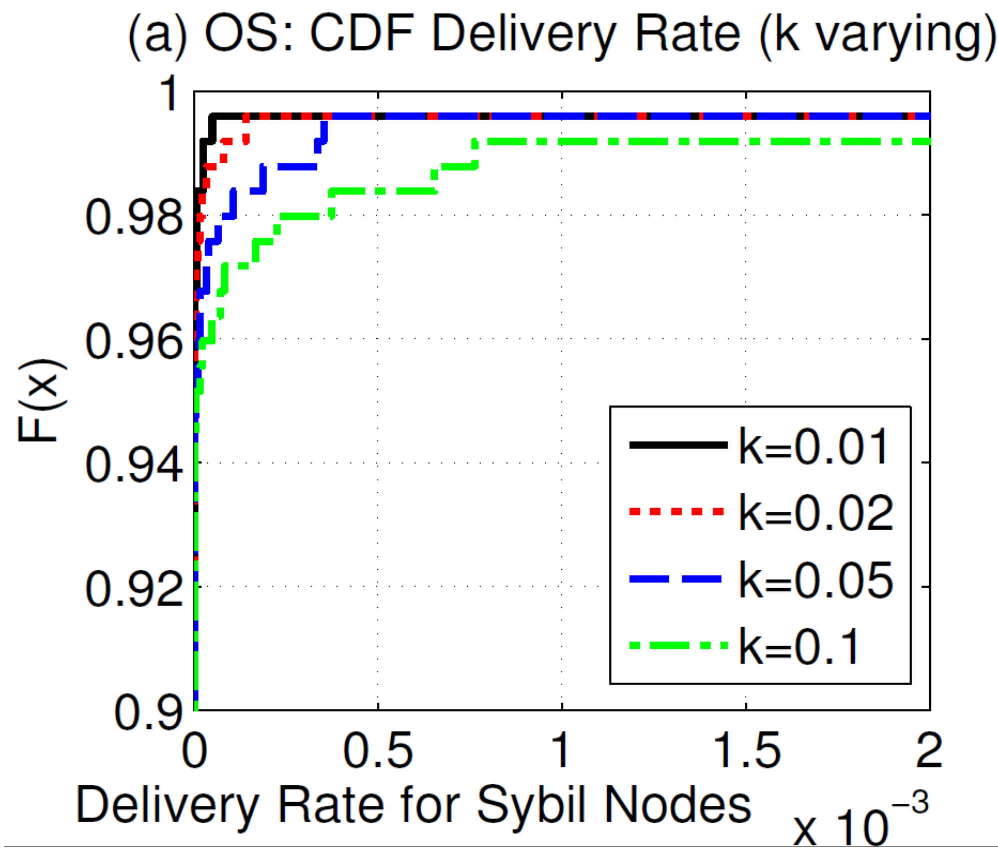
ST Approaches – Example: Ostra

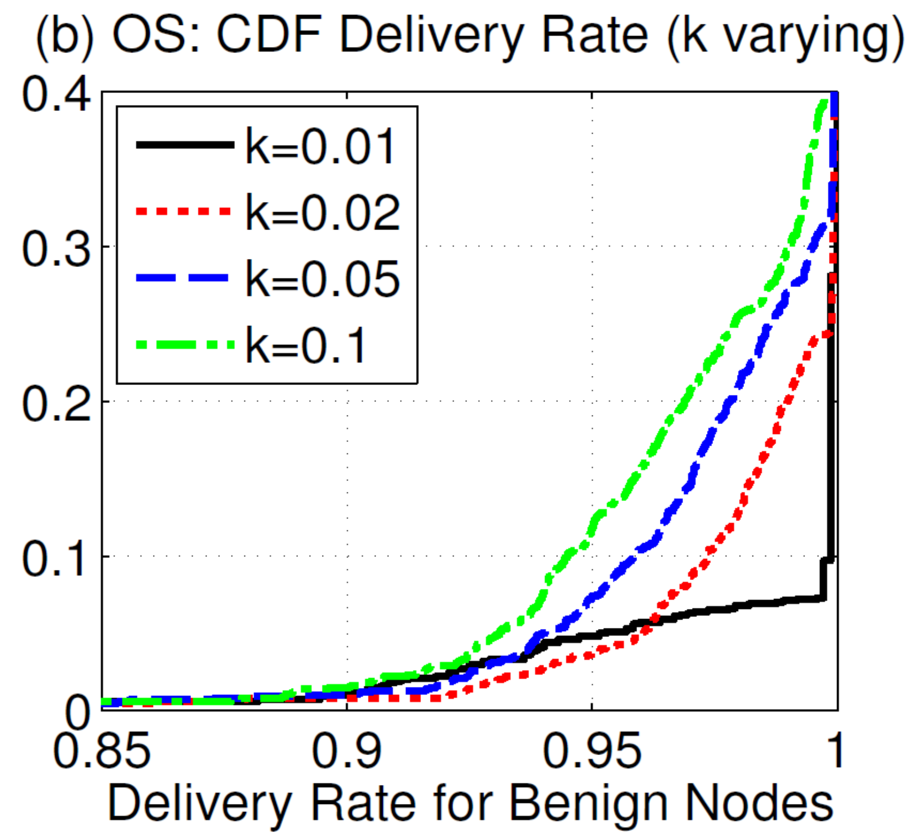
- Dependency on attack edges
  - Amount of spam grows proportionally to number of attack edges
- But there's more:
  - Spam sent along critical edges also affects benign nodes!
  - Communities may be blocked from sending to outside!



Evaluation: Ostra Performance

- Here: k = overall ratio of attack edges in network





ST Approaches - Overview

- ST approaches have the same general working principle, but more specific weaknesses
- Reason: Designed for a specific application
  - e.g., in SumUp (a vote collection scheme):

    An intelligent voting strategy can lead to attackers outvote honest users

Summary

- Previous assumptions for Sybil Defenses do not hold anymore

- We reveal severe weaknesses in all recent Sybil Defenses revealed by qualitative and quantitative analysis
  - Low distinguishing ability of solutions
  - In SD approaches, mostly ***1 or 2 attack edges*** are enough
  - In ST approaches, issues are more specific, but still severe

What do future OSN-based approaches need?

- Use meta-data of relations in addition to graph structure itself
  - Intensity of the relation (e.g., message frequency)
    - But: High false positive rate?
  - Lifetime of a user's relations (i.e., a node is suspicious if a lot of its relations are short-lived)

- Challenge: How to get a data set that would provide such info for testing the approach and verifying it?

Thank You! Any Questions?

[1] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In *WWW ’09*. ACM, 2009.

[2] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai. Uncovering social network sybils in the wild. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, IMC ’11, pages 259–268, New York, NY, USA, 2011. ACM.

[3] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC ’11, pages 93– 102, New York, NY, USA, 2011. ACM.

[4] D. Irani, M. Balduzzi, D. Balzarotti, E. Kirda, and C. Pu. Reverse social engineering attacks in online social networks. In *Proceedings of the 8th* international conference on Detection of intrusions and malware, and vulnerability *assessment*, DIMVA’11, pages 55–74, Berlin, Heidelberg, 2011.

[5] V. Sridharan, V. Shankar, and M. Gupta. Twitter Games: How Successful Spammers Pick Targets, to appear in ACSAC'12

This work has been partially supported by the NSF (grant no. CNS-0644434 and CNS-1118101).

References

- [1] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In *WWW '09*. ACM, 2009.
- [2] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai. Uncovering social network sybils in the wild. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, IMC '11, pages 259–268, New York, NY, USA, 2011. ACM.
- [3] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC '11, pages 93– 102, New York, NY, USA, 2011. ACM.
- [4] D. Irani, M. Balduzzi, D. Balzarotti, E. Kirda, and C. Pu. Reverse social engineering attacks in online social networks. In *Proceedings of the 8th* international conference on Detection of intrusions and malware, and vulnerability *assessment*, DIMVA'11, pages 55–74, Berlin, Heidelberg, 2011.
- [5] V. Sridharan, V. Shankar, and M. Gupta. Twitter Games: How Successful Spammers Pick Targets, to appear in ACSAC'12