Research Methods

Prof. William Enck
Reading papers ...

- What is the purpose of reading papers?
- How do you read papers?
Understanding what you read

• Things you should be getting out of a paper
  ‣ What is the central idea proposed/explored in the paper?
    • Abstract
    • Introduction
    • Conclusions

  These are the best areas to find an overview of the contribution

  ‣ How does this work fit into others in the area?
    • Related work - often a separate section, sometimes not, every paper should detail the relevant literature. Papers that do not do this or do a superficial job are almost sure to be bad ones.

    • An informed reader should be able to read the related work and understand the basic approaches in the area, and how they differ from the present work.
Understanding what you read (cont.)

• What scientific devices are the authors using to communicate their point?

• **Methodology** - this is how they evaluate their solution.
  
  ‣ Theoretical papers typically validate a model using mathematical arguments (e.g., proofs)
  
  ‣ Experimental papers evaluate results based on test apparatus (e.g., measurements, data mining, synthetic workload simulation, trace-based simulation).

• **Empirical** research evaluates by measurement.

  ‣ Some papers have no evaluation at all, but argue the merits of the solution in prose (e.g., position papers)
Understanding what you read (cont.)

• What do the authors claim?
  
  ‣ **Results** - statement of new scientific discovery.
  
  • Typically some abbreviated form of the results will be present in the abstract, introduction, and/or conclusions.
  
  • **Note**: just because a result was accepted into a conference or journal does necessarily not mean that it is true. Always be circumspect.

• What should you remember about this paper?

  ‣ **Take away** - what general lesson or fact should you take away from the paper.

  ‣ Note that really good papers will have take-aways that are more general than the paper topic.
Summarize Thompson Article

- Contribution
- Motivation
- Related work
- Methodology
- Results
- Take away
A Sample Summary

• **Contribution:** Ken Thompson shows how hard it is to trust the security of software in this paper. He describes an approach whereby he can embed a Trojan horse in a compiler that can insert malicious code on a trigger (e.g., recognizing a login program).

• **Motivation:** People need to recognize the security limitations of programming.

• **Related Work:** This approach is an example of a Trojan horse program. A Trojan horse is a program that serves a legitimate purpose on the surface, but includes malicious code that will be executed with it. Examples include the Sony/BMG rootkit: the program provided music legitimately, but also installed spyware.

• **Methodology:** The approach works by generating a malicious binary that is used to compile compilers. Since the compiler code looks OK and the malice is in the binary compiler compiler, it is difficult to detect.

• **Results:** The system identifies construction of login programs and miscompiles the command to accept a particular password known to the attacker.

• **Take away:** Thompson states the “obvious” moral that “you cannot trust code that you did not totally create yourself.” We all depend on code, but constructing a basis for trusting it is very hard, even today.
Reading a paper

• Everyone has a different way of reading a paper.

• Here are some guidelines I use:
  ‣ Always have a copy to mark-up. Your margin notes will serve as invaluable sign-posts when you come back to the paper (e.g., “here is the experimental setup” or “main result described here”)
  ‣ After reading, write a summary of the paper containing answers to the questions in the preceding slides. If you can’t answer (at least at a high level) these questions without referring to the paper, it may be worth scanning again.

• Over the semester, try different strategies for reading papers (e.g., Honeyman approach) and see which one is the most effective for you.
Reading a (systems) security paper

• What is the security model?
  ‣ Who are the participants and adversaries
  ‣ What are the assumptions of trust (trust model)
  ‣ What are the relevant risks/threats

• What are the constraints?
  ‣ What are the practical limitations of the environment
  ‣ To what degree are the participants available

• What is the solution?
  ‣ How are the threats reasonably addressed
  ‣ How do they evaluate the solution

• What is the take away?
  ‣ key idea/design, e.g., generalization (not solely engineering)

• Hint: I will ask these questions when evaluating course project.
Why write a paper?

• There are many reasons to write a paper:
  ‣ Articulate a new idea, thought, or observation ...
  ‣ Document your research ...
  ‣ Talk about new (observed) phenomenon ....
  ‣ Advance your career ...
  ‣ Because you have to ...

• **Reality**: publication is the coin of the realm in science, failure to do this successfully will lead to failure. You have to be effective at this to be a good (a) graduate student, (b) faculty member, or [sometimes] (c) researcher in professional research laboratory (IBM/AT&T/MSR)
Where to publish?

• Venues for publication:
  ‣ Tech report
  ‣ Workshop
  ‣ Conference
  ‣ Journal
  ‣ Book

• Often your work will work through these from *preliminary* to *archival* versions of the work, sometimes branching or joining.

• **Book**: less frequent, more work.
Publication Tiers

• Not all publication venues are valued the same. Publication “tiers” tell the story

• 1st tier - IEEE S&P, USENIX Sec, CCS, \textit{TISSEC, JCS}
  ‣ 1.5 NDSS

• 2nd tier - ACSAC, ACNS, ESORICS, CSF, RAID, \textit{TOIT}

• 3rd tier - SecureComm, ICISS

• 4th tier - HICS
  ‣ SClgen (WMSCI 2005)
• The *editor-in-chief* (EIC) receives the papers as they are submitted.

• The papers are assigned to *associate editors* for handling.

• *Anonymous reviewers* rate the paper:
  - Accept without changes
  - Minor revision
  - Major revision
  - Reject
The **PC Chair** is the person who marshals the reviewing and decisions of a conference. This is different than the **general chair**.

**PC members** review, rate and discuss the paper, then vote on which ones are accepted.

The **acceptance rate** is the ratio of accepted to submitted papers.
Paper evaluation

• A paper is evaluated on
  ‣ Novelty
  ‣ Correctness
  ‣ Impact
  ‣ Presentation
  ‣ Relevance
  ‣ “hotness”
What is research?

• Which activities are research?
  ‣ Designing a new protocol?
  ‣ Building an implementation of a protocol?
  ‣ Measuring the cost of the protocol?
  ‣ Formally evaluating the correctness of a protocol?
  ‣ Developing methods of implementing, evaluation a protocol?
What is not research?

- Arguing the quality of a protocol?
- Arguing the appropriateness of a protocol?
- Surveying a field?
- Illustrating a limitation of a common practice or system?
A cynical definition:

• That which counts on your vita … is research.

• The hardest thing about a PhD is figuring out what “research” is …
Research vs. engineering

- Novelty …
- Importance … (sort of)
- Discovering a new fact or idea

- Engineering is often harder than research
- One must be careful to understand the difference
Research vs. Opinion

• Arguing a position is not research unless it uncovers some new thought or methodological device
  ‣ Difference is subtle

• Experts will often produce manifesto about an area
    – The key here is that they are experts and have the bona fides to make some an argument
    – This is not research
Why is there so much bad research?

- Most papers (90%+) I encounter are bad --- for one or more of the following reasons. The authors …
  … don’t formulate the problem well (or at all).
  … don’t motivate the problem well (or at all).
  … address an unimportant or moot problem.
  … are not familiar with the breadth or depth of the area.
  … do not discuss important related work.
  … don’t realize the problem has been solved (or at least better addressed).
  … don’t have a coherent solution or it does not solve the problem.
  … don’t have a coherent or appropriate methodology.
  … don’t apply the methodology well.
  … don’t draw the correct conclusions from the results.
  … don’t present the work well enough to be understandable.
  … don’t articulate the take away.

- Any paper failing to do any of these things is a failure.
Security Research

• Almost as diverse as computer science itself
  ‣ Systems design
  ‣ Formal analysis
  ‣ Programming languages
  ‣ Hardware design
  ‣ Software engineering
  ‣ Human computer interfaces
  ‣ Networking, …

• Some are specific to security
  ‣ Cryptography
  ‣ Security protocol design
  ‣ Security Policy …
Idea Formulation

• The essential part of successful research is picking good problems and solutions?

• Q: How do you do this?
Idea Formulation (cont.)

- Good approaches to finding ideas:
  - First, read several papers (make sure they are good ones) in a particular area.
    - If this is a new topic area, you must become familiar with the problems, solutions, and terminology of the community.
  - Then ask the following questions (write down answers)
    - What are the problems that this area addresses?
    - What are the methodological tools that people bring to bear in addressing problems in this area?
    - How is the field evolving?
    - How do your set of skills apply to the problems being addressed?
    - How are expected changes in the larger computer science community going to affect the known problems and solutions?

- Paper: “Patch on Demand” Saves Even More Time?
Idea Formulation - LISTING

• Do the following exercises:
  ‣ (5 minutes) Listing: make a quick list of 1-5 word phrases that would be used by/related to/observance of the field and problems and solutions
    • This is **not** an outline, there is no ordering to the list
    • Use your imagination
    • Creativity is the essence of this exercise (**don’t overthink**)
    • Some of list will be nonsense, do not filter thoughts
  ‣ Example: if I were looking at a paper about firewalls, I might come up with the following (just a start):
    • policy validation, distributed firewalls, bad for detecting viruses, …
    • Of course, this is general, should contain thoughts more specific to paper content,
    • e.g., better algorithm than Bob (the author) -- use graph theory
storage provenance, network provenance, tracking information as it goes between systems in the cloud, state of systems when creating data, processing data, sending data to the next stage, pipelines of information flow, pipelines in SCADA systems, relation of provenance to real world workflows, real world workflows vs workflows of information between applications, how isolated are applications in their data use?, many phone applications are isolated, but communicate with cloud servers, are smartphone apps producers or consumers of information?, does this related to provenance anymore? healthcare workers use smartphones rather frequently, can geographic location be used as a provenance source in a phone-cloud system? location and provenance are both sometimes used for access control.
Using the results

• Examine closely the contents -- they will tell a story
  find singletons or clusters or phrases and see if they
  provide some new angle on a problem or issue

• For example, I choose: geographic location be used as
  a provenance source

• Which leads the following idea:
  ‣ Q: In what environments can location provenance be used?
  ‣ Q: What real-world analogies are there?
  ‣ Only read something that was written in a similar spacial/
    provenance context
    • Paper: “Situational Memory Recall for Access Control Policy”
Now do it.

- 5 minutes - scan paper
- 1 minute - free writing
- 2 minute - develop idea
Papers

• Abstract
• Introduction
• Background/Motivation
• Solution
• Experiment/Evaluation
• Discussion
• Conclusions
Abstracts (Purpose?)

• Communicate the content of a paper in enough depth that the reader can broadly understand its *scope* and *contribution* in 30 seconds of reading.

• I.e., answer the following questions …
  ‣ What is this paper about?
  ‣ What can I expect to learn from it?
  ‣ What did you find?
  ‣ How did you find it?

• Often least effort, but it is one thing you are guaranteed the reviewers and reader will read in entirety.
How do you write an abstract?

• Interestingly, the abstract you write when you start is often not the abstract you would write when you are finished.
  ‣ Sales-pitch for the paper, and it will change as you write.
  ‣ Implication: *Always* go back and rewrite the abstract last.
  ‣ You can “lose” the acceptance in abstract, but not “win” it.
  ‣ Careful of grammar, spelling, style. -- reviewers will develop an impression of the work based on a quick read of the abstract.
  ‣ Make sure it actually is concrete about work--overly broad, vague or vacuous abstracts are severely punished.

• Q: OK, what goes in it?
One method: 6 points

1. **Area** - what is the basic area about?
2. **Problem** - what problem are we trying to solve?
3. **Solution** - how do you address that problem?
4. **Methodology** - how do you validate/evaluate solution?
5. **Results** - what does that evaluation show?
6. **Take-away** - what is the broader lesson/result?

• **Note**: this must be stand-alone.
Protection systems exist to prevent the leakage or corruption of system and user data. Traditional discretionary access control mechanisms do not differentiate between a user's running applications and hence provide no means of preventing one application from exploiting another's data. Because commercial mandatory access control mechanisms, such as SELinux and AppArmor, aim to protect system files, they can do little to prevent similar misuse of user data. This paper presents the PinUP access control overlay which extends filesystem protections by limiting the set of user applications that can access the user's high-value files. We describe our model, architecture, and Linux implementation, evaluate run-time costs, and detail use-cases illustrating the power and utility of the augmented policy. Our performance experiments show that all costs are nominal, with a maximum observed delay of 40 milliseconds occurring at application startup and a few tens of microseconds at each access check. In this, we provide efficient application-oriented access controls that avoid inter-application misuse of user data.

(Area, Problem, Solution, Methodology, Results, Take-Away)
Protection systems exist to prevent the leakage or corruption of system and user data. Traditional discretionary access control mechanisms do not differentiate between a user's running applications and hence provide no means of preventing one application from exploiting another's data. Because commercial mandatory access control mechanisms, such as SELinux and AppArmor, aim to protect system files, they can do little to prevent similar misuse of user data. This paper presents the PinUP access control overlay which extends filesystem protections by limiting the set of user applications that can access the user's high-value files. We describe our model, architecture, and Linux implementation, evaluate run-time costs, and detail use-cases illustrating the power and utility of the augmented policy. Our performance experiments show that all costs are nominal, with a maximum observed delay of 40 milliseconds occurring at application startup and a few tens of microseconds at each access check. In this, we provide efficient application-oriented access controls that avoid inter-application misuse of user data.

1. Area
2. Problem
3. Solution
4. Methodology
5. Results
6. Take-Away
Your Turn ...

1. Area
2. Problem
3. Solution
4. Methodology
5. Results
6. Take-Away

• **Paper**: the one you read for the idea generation.
Contribution

• Somewhere in the introduction, you have to say what the contribution of the paper is ...
  ‣ typically, this is stated rather explicitly in a single declarative paragraph
  ‣ most papers repeat this in conclusions
  ‣ If you are missing this, then the paper will be confusing

• Q: What should it contain?
“This paper considers how the operational characteristics of BGP can be exploited to close the security infrastructure cost/security model gap. The central observation driving this work is that the vast majority of ASes offer few distinct paths for a prefix, and that those paths are largely static. We confirm this through a study of path stability. We study the 40 RouteViews listening points, and found that in the average case, less than 2% of prefixes were advertised using more than 10 paths, and less than 0.06% were advertised with more than 20 paths during a single month.”

• Broad approach leads to
• Key observation leads to
• Main result leads to
Importance

• This should be a statement of why you should care about the problem or solution
  ‣ Think very carefully, this is different that motivation (why?)
  ‣ What is the right level to state your contribution?
  ‣ You should be careful not to overstate the importance
  ‣ You should be careful not to understate the importance
“The Six-Hats method may well be the most important change in human thinking for the last 2300 years. That may seem a rather exaggerated claim, but the evidence is beginning to point that way.”

- Edward De Bono (Six Thinking Hats)
• The approach has been validated through several case studies of attacks ... Some of these attacks were not detectable ...

- Anonymous
Practice

• We are going to write a contribution importance paragraph for your paper. However, you must do it now from scratch:

• Caveat: it *must* be structured as follows:
  1. broad approach
  2. key observation
  3. main result
  4. importance
Related Work

• Q: What is the point of writing a related work section?

• A: To establish ....

  ‣ Need for work.
    • Why previous works don’t get it done ...
    • The limitations of past work ...
  
  ‣ Mastery over area.
    • Established bona fides ...
  
  ‣ Relationship to other scientific areas.
    • How relates to bodies of other works ...
  
  ‣ Others?
How?

• Common, wrong, way to write a paper.
  ‣ Algorithms a, b, c, and d. have been done.
    • A is good because of Blah, bad because of Duh.
    • B is good because of Blah’, bad because of Duh’.
    • C is good because of Blah”, bad because of Duh”.
    • ..... 

• A laundry list with no introspection about field in which it exists.
Narrative

- Tell a story about the field in which it exists. It should try to organize in such a way as you can see how the work evolves from start to finish.
- Ideally, ends with the conclusion that the present work is needed.
- Example:
  - (para 1) The early algorithms, A and B, sought to solve the requirements X, Y, and Z by bit-twiddling. However such approaches eventually lead to blah and duh, to differing degrees.
  - (para 2) C, D, and other also tried to solve the same problem, but failed ...
  - (para 3) Ultimately, the authors were not attractive enough to solve the problem, so it was left to me.
Related work (summarized)

• A good related work section should include works …
  ‣ If they address the central problem
  ‣ If they address a related problem
  ‣ If they identified the problem
  ‣ If they use the same methodology for a similar problem
  ‣ If your work was inspired by them

• It should be a narrative about the field, its logical relatives, the problems it faces, advances and failures, and motivating articles.
  ‣ Show how the body of work holds together in some philosophical or technological way
  ‣ Demonstrate mastery of subject matter to establish credentials for paper (often a fatality if done wrong)
Experiments

- Q: What is the purpose of experiments in a scholarly research paper?

- A: It depends on the paper.
Experimental Analysis

• Experiments are used to
  ‣ Demonstrate/explain some unknown phenomena
  ‣ Explore the tradeoffs between solutions
  ‣ Illustrate the accuracy of models
  ‣ Demonstrate implementation
  ‣ …
The hypothesis

• All experimental methodology must start with a hypothesis (assumption in ancient Greek)
  ‣ A statement of truth that you would like to validate
  ‣ Sometimes it may lead to unanswerable questions

• Examples:
  ‣ The earth is round
  ‣ NC State is in North Carolina
  ‣ Signal strength can accurately be estimated by distance
  ‣ Gravity is a function of mass
  ‣ People like ice cream
  ‣ NCSU has a good graduate program
Hypothesis Testing

• The experimental apparatus should lead to some answer to the question

• … put another way, questions dictate experiments

• Caution: don’t be a hammer user
Experimental Approaches

• Simulation
  ‣ Model using simplified or abstracted features (e.g., worm experiments)

• Emulation
  ‣ Model the behavior in detail (e.g., VMs)

• Measurement
  ‣ Measure real phenomena in real environment (e.g., address measurement)
Deciding on experimental methodology

• Lets try it …

1. The earth is round
2. NC State is in North Carolina
3. Signal strength can accurately be estimated by distance
4. Gravity is a function of mass
5. People like ice cream
6. NCSU has a good graduate program
• All experiments should be documented in sufficient detail to be *repeatable*
  ‣ This is the foundation of good science
  ‣ A key part of scientific ethics is data management

• All data/apparatus should be made public
  ‣ In some cases, this might not be possible
  ‣ Open source tools are useful (e.g., CQual) -- good ones are publishable
Writing up results

• Experimental setup
  ‣ give all the grungy details about environment (repeatability), source data, tools
  ‣ Broad definition of experimental areas

• Experiments
  ‣ Generally, I like to create subsections for every major question/answer
  ‣ Each section should “signpost” the question, then explain the experiment
  ‣ Each major feature should be given in a single paragraph
Conclusions

• A conclusion is a way to not end abruptly.

• Only a fraction of people will read it closely.
  ‣ No new information
  ‣ Just a restatement (in past tense) of the results
  ‣ Highlight the take-away

• Some add future work
  ‣ Only if you plan to move forward.
  ‣ Be careful.
Closing Note: Authorship

• This is the *most dangerous part of publishing*. This has led to the most serious rifts in the profession …
  ‣ Make sure that anyone involved knows the policy (what one needs to do to be an author) the expectations and the repercussions of not participating as expected.

• Ordering matters in some fields (systems), not in others (math).
• Make sure everything is clear to everyone before getting started.
  ‣ I have seen best friends never speak to each other again.
  ‣ A paper is never worth that kind of heartache, but people will surprise you.
  ‣ Do you have a policy and what is it?
... a few more things

• Putting pen to papers takes some planning.

• How do you start writing?
  ‣ Brain dump
  ‣ Outline
  ‣ Collect lots of graphs