accenture

Confidentiality Protection in Crowdsourcing

Simran Saxena

Dr. Ponnurangam Kumaraguru (Chair)
Dr. Alpana Dubey (Co-chair)

linkedin/in/simransaxena

@simran_s21

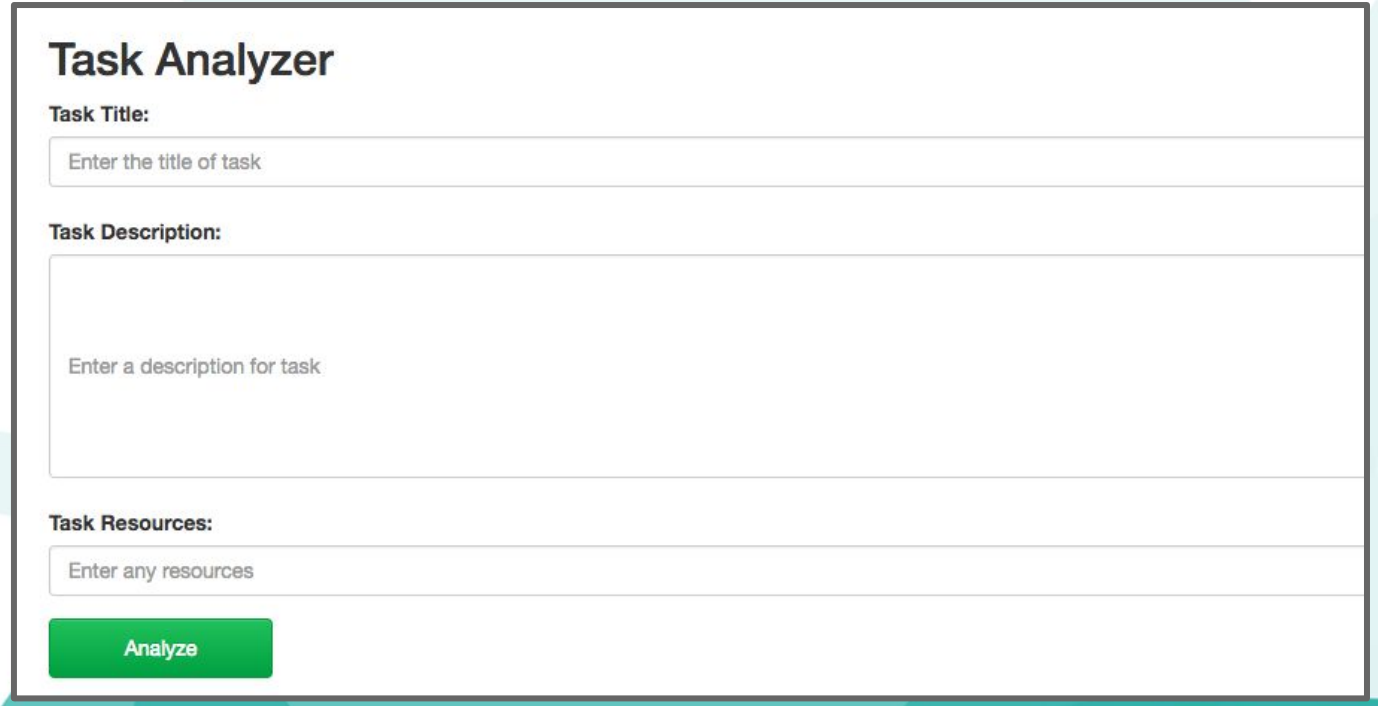
fb.com/simran.s21

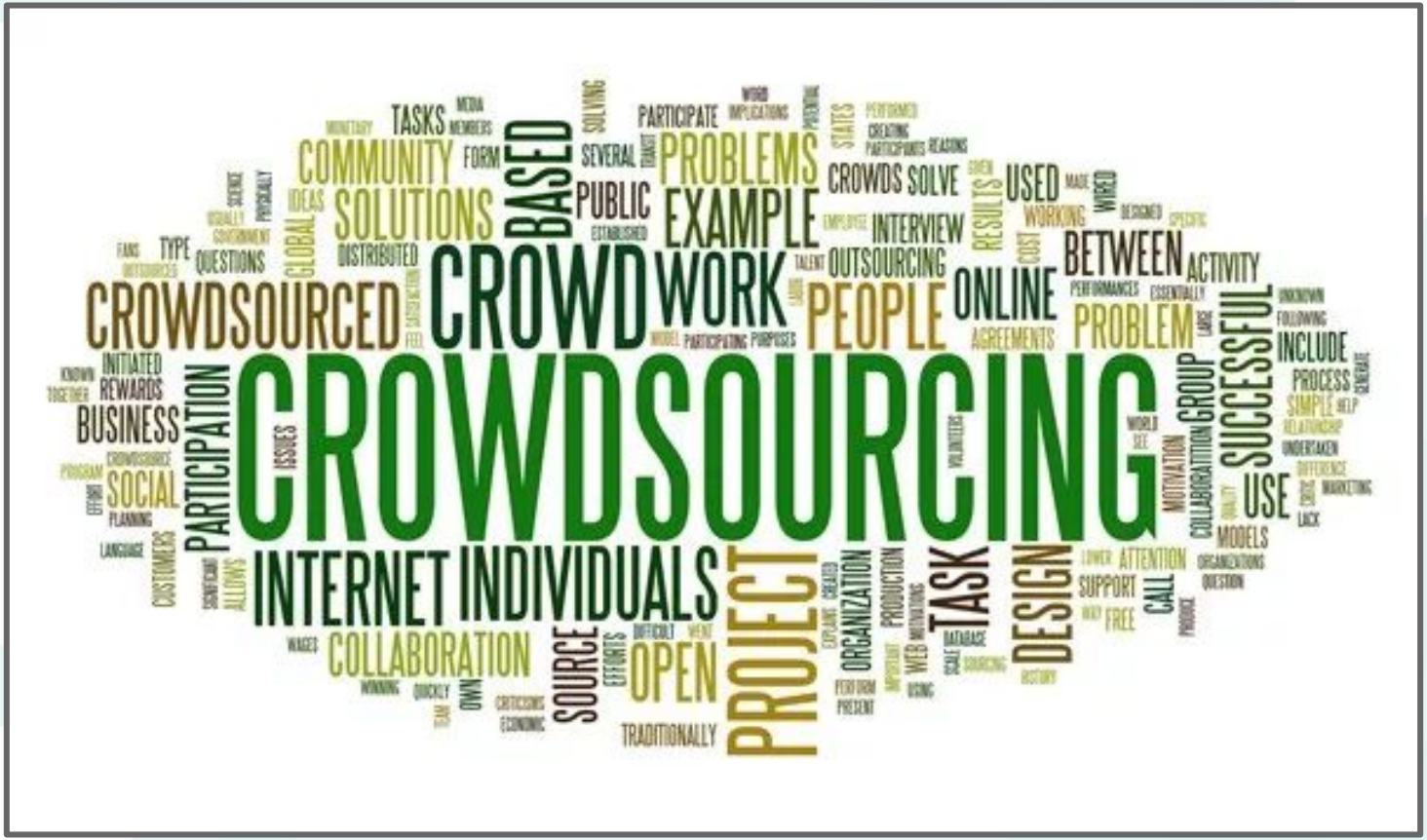
Thesis Committee

- Dr. Arun Balaji Buduru, IIIT Delhi
- Dr. Niharika Sachdeva, InfoEdge
- Dr. Alpana Dubey, Accenture Technology Labs
- Dr. Ponnurangam Kumaraguru, IIIT Delhi

Demo

- Proof of Concept




CROWDSOURCING
CROWD
CROWDSOURCED
WORK
PEOPLE
ONLINE
BASED
EXAMPLE
PROBLEMS
PUBLIC
COMMUNITY
SOLUTIONS
TASKS
MEDIA
MEMBERS
FORM
SEVERAL
SOLVING
PARTICIPATE
WORD
IMPLICATIONS
STATS
PERFORMED
CREATING
PARTICIPANTS
REASONS
CROWDS
SOLVE
USED
MADE
WIRED
INTERVIEW
RESULTS
WORKING
DESIGNED
SPECIFIC
BETWEEN
ACTIVITY
OUTSOURCING
TALENT
COST
ESTABLISHED
DISTRIBUTED
QUESTIONS
TYPE
FANS
OUTSOURCED
GLOBAL
IDEAS
SCIENCE
PHYSICALLY
GOVERNMENT
MONETARY
PROBLEM
AGREEMENTS
PERFORMANCES
ESSENTIALLY
UNKNOWN
FOLLOWING
INCLUDE
PROCESS
SIMPLE
HELP
GENERATE
RELATIONSHIP
UNDERTAKEN
DIFFERENCE
MARKETING
LACK
SUCCESSFUL
GROUP
USE
MODELS
QUALITY
COLLABORATITION
MOTIVATION
WORLD
SEE
ORGANIZATIONS
QUESTION
ATTENTION
LOWER
SUPPORT
CALL
FREE
WAY
PRODUCE
MODEL
PARTICIPATING
PURPOSES
FEEL
INITIATED
KNOWN
TOGETHER
REWARDS
BUSINESS
PARTICIPATION
ISSUES
CROWDSOURCE
PROGRAM
SOCIAL
EFFORT
PLANNING
LANGUAGE
CUSTOMERS
ALLOWS
SIGNIFICANT
INTERNET
INDIVIDUALS
PROJECT
ORGANIZATION
PRODUCTION
TASK
DESIGN
VOLUNTEERS
CREATED
EXPLAINS
WAGES
COLLABORATION
WINNING
QUICKLY
OWN
TEAM
CRITICISMS
ECONOMIC
SOURCE
EFFORTS
DIFFICULT
WENT
OPEN
TRADITIONALLY
PERFORM
PRESENT
IMPORTANT
WEB
USING
MOTIVATIONS
SCALE
DATABASE
SOURCING
HISTORY

Outline

- Research Motivation
- Research Aim
- Crowdsourcing at the level of organizations
- Survey to understand confidentiality
- Unboxing a typical crowdsourcing task
- Understanding conversations between workers and task posters
- Protecting confidentiality loss in crowdsourcing
- Conclusion

Shift Towards a Gig-Economy: Benefits

Cost Effective

Access to a diverse talent pool

Faster delivery

> *The fact that there is this quality of engineers all over the world, it really made me re-think how I resource the projects.*
>
> **Marc Kocher**
> Sr. Manager, Engineering

Ref: http://isroset.org/pub_paper/IJSRCSE/2-IJSRCSE-0315.pdf

Benefits of Crowdsourcing

Cheaper and faster
60% of the times

Reduces cost by
30 - 80%

Reduced ***time-to-market***
because of Parallelism

Ref: 1, 2

Happiness and Satisfaction Levels

74%

72%

95%

Ref: https://fieldnation.com/wp-content/uploads/2017/03/The_2016_Field_Nation_Freelancer_Study_R1V1__1_-2.pdf

Crowdsourcing as the Next Big Revolution

By ***2020***, freelancers will form ***43%*** of the workforce in USA

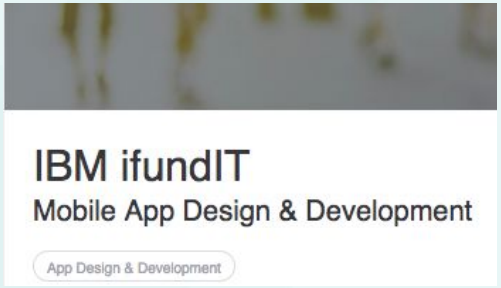
Crowdsourcing in Organizations

Many popular companies make use of Crowdsourcing

Back to Customer Stories

ebay

IT requests cost time and money. eBay wanted a sleek mobile application to optimize the worker experience and it needed to be dead simple to use. See how crowdsourcing delivered.



Crowdsourcing and Software Development

Coding

Data Analysis

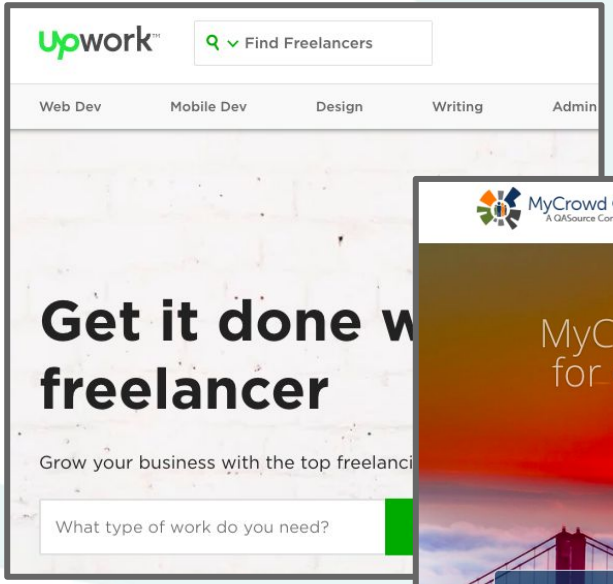
Web Development

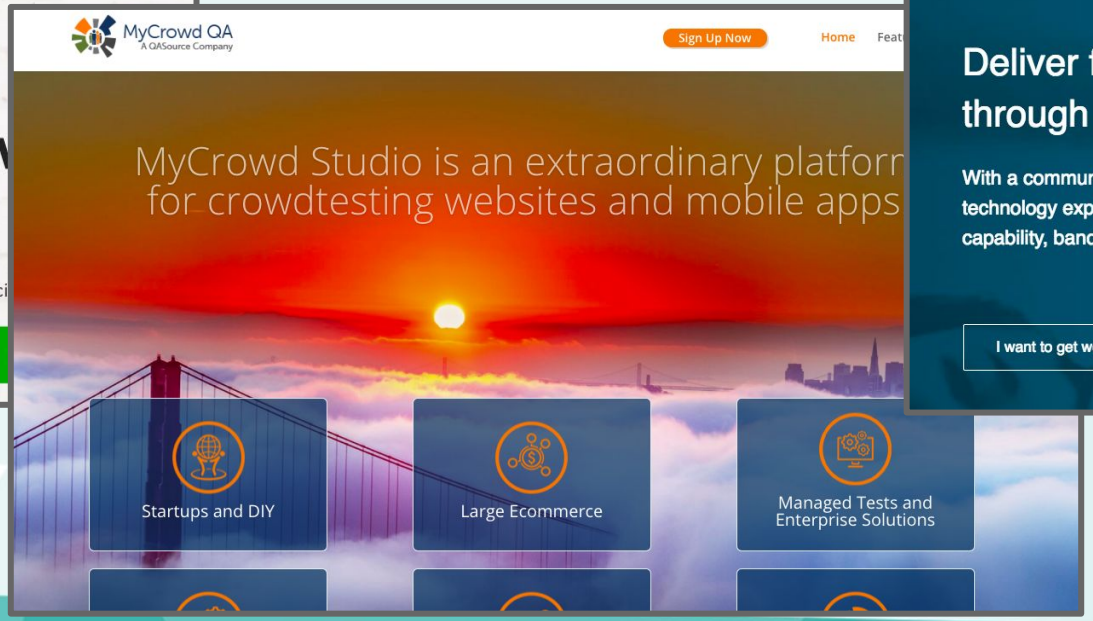
Designing, Prototyping

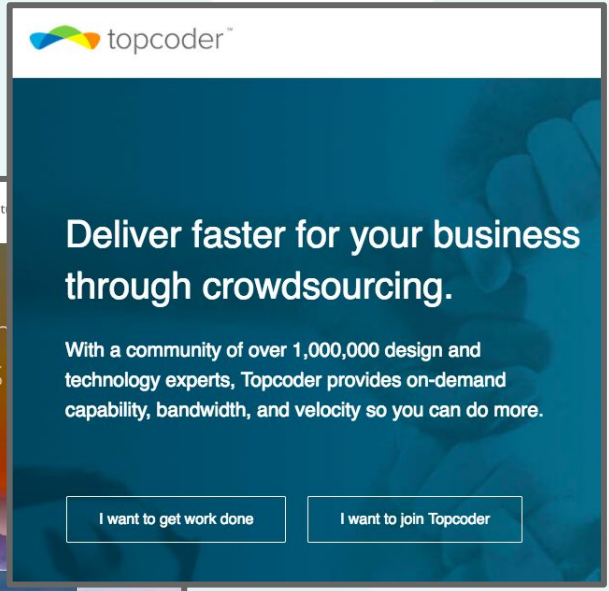
Testing

Mobile Development

Crowdsourcing Platforms







And many more..

Stakeholders in Crowdsourcing

Task Poster -> ***tp***

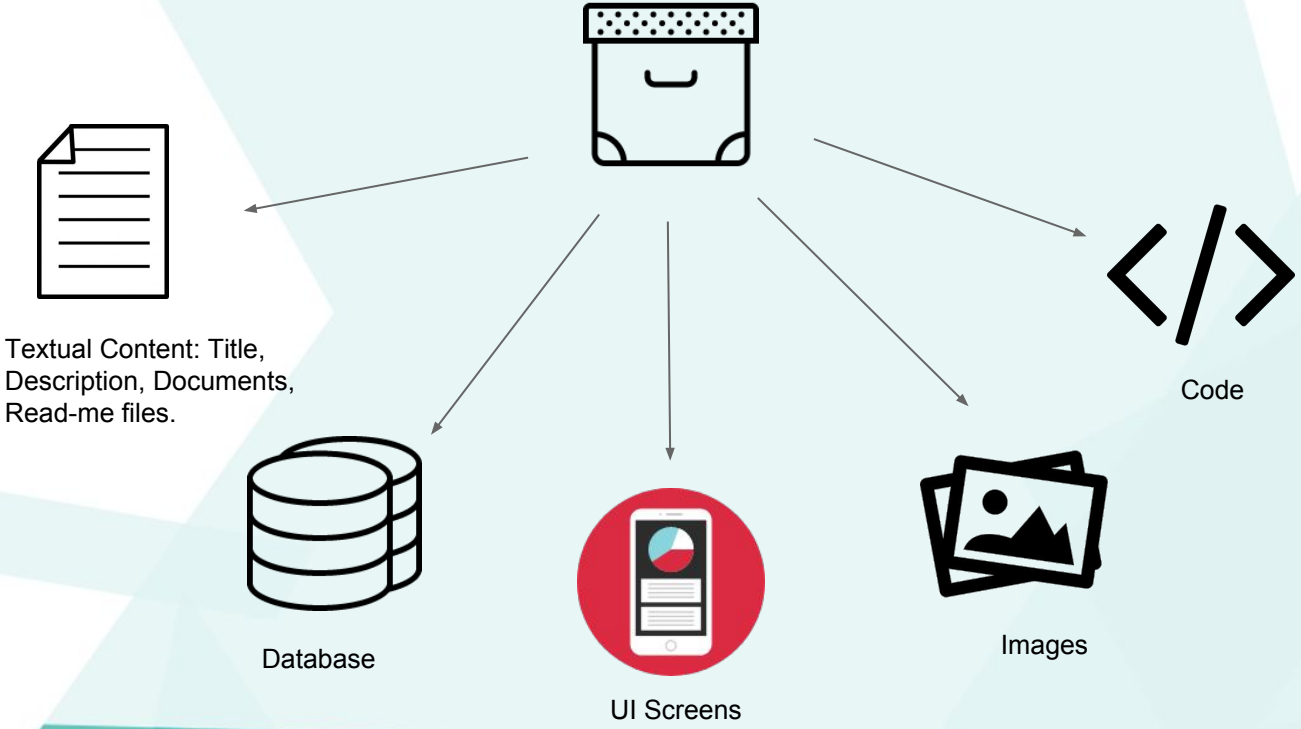
Worker -> ***w***

Company/Organization -> ***c***

Components of a Crowdsourcing Task

- ~ 20 attributes per task
  - Task ID
  - Task Title
  - Task Description
  - Task Category
  - Task Type
  - Task Workload
  - ….

Resources shared commonly



Sample Software Development Task [1]

Password generator with barcode

Desktop Software Development

Hi,

I'm looking for a JAVA software that can generate strong password with barcodes.

The software should allow me to set:

password length (max 40)
options to include symbols i.e @#$%
options to include numbers
options to include lowercase characters and upper case characters.

And this software cannot produce the same password.

Project Stage: N/A

Operating systems: Windows

Hours to be determined
Hourly

Less than 1 week
Project Length

$$ **Intermediate Level**
I am looking for a mix of experience and value

June 21, 2017
Start Date

Post a job like this

Submit a proposal

About the Client

★★★★★ (5.00) 5 reviews

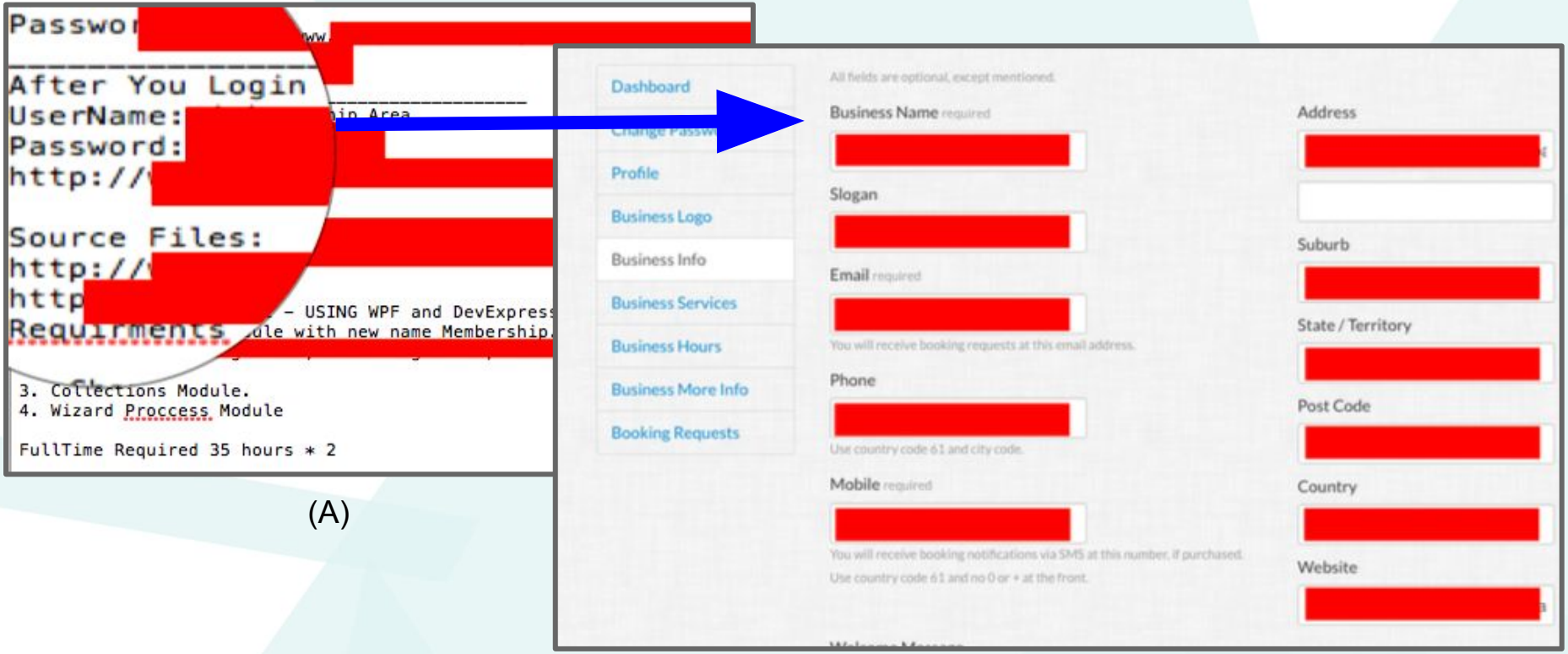
Canada
Montreal 05:46 AM

22 Jobs Posted
60% Hire Rate, 1 Open Job

$5k+ Total Spent
13 Hires, 0 Active

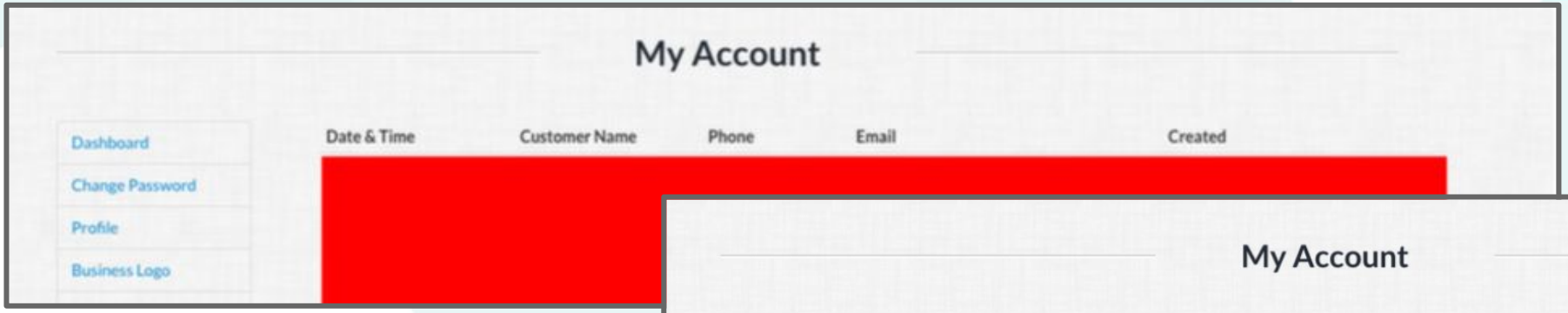
$12.93/hr Avg Hourly Rate Paid
74 Hours

Why is it a problem?

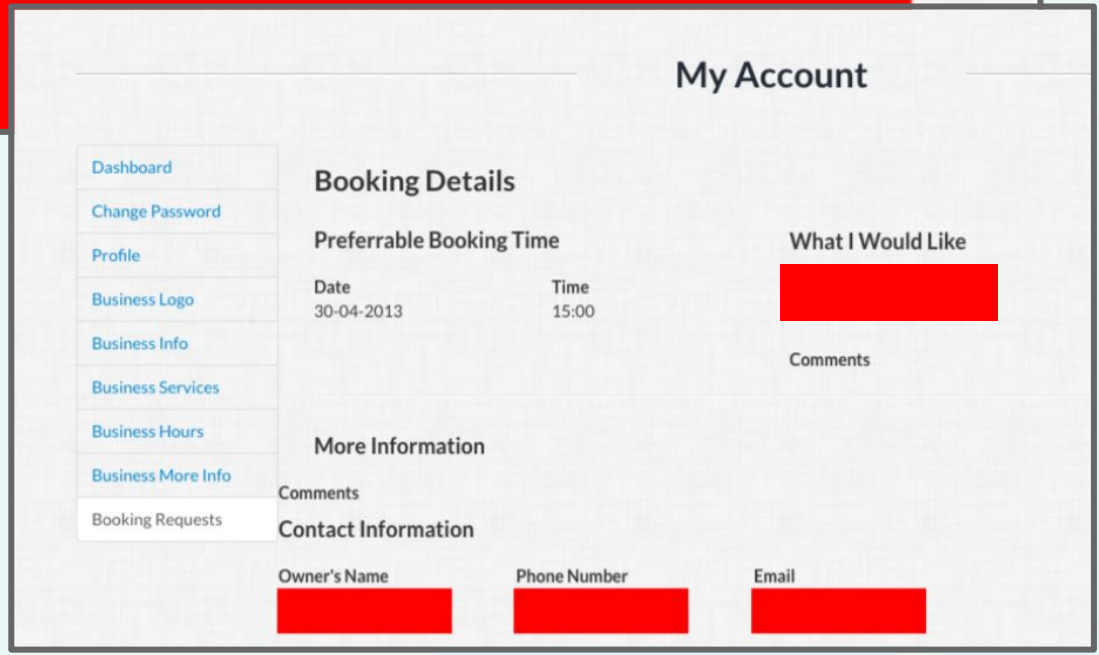


(A)

(B)



(A)



(B)

Exposed data of the Customers:

- Name
- Address
- Email ID
- Phone number
- Details of availed services
  - Date
  - Time
  - Cost

Why is it a problem?

The above form will then be submitted to our server side PHP file. This PHP page will need to make the charge with code similar to the following sample:

```
// sk key in production
your keys here https://[redacted]/account
setApiKey("sk_test_BQshsh3PolfQ2");
the credit card details submi the form
... = $_POST['[redacted]']
...
...
...
```

(A)

My $2375 Amazon EC2 Mistake ☞

Jan 7th 2015 in Commentary

A word of warning: Know what your modules/extensions/pods/plugins are doing, **especially** if they use any of your credentials.

I'm actually surprised that this actually was up that long. I accidentally did this once and Amazon was on the phone with me 10 mins later.

Turns out through the S3 API you can actually spin up EC2 instances, and my key had been spotted by a bot that continually searches GitHub for API keys. Amazon AWS customer support informed me this happens a lot recently, hackers have created an algorithm that searches GitHub 24 hours per day for API keys... Once it finds one it spins up max instances of EC2 servers to farm itself bitcoins...

Boom! A $2375 bill in the morning. Just for trying to learn rails.

(B)

Why is it a problem?

Name

[redacted] Testing Workbench Solution - TCS Presentat

Consolidated [redacted] Menus_Details V1.xlsx

Consolidated [redacted] Menus_Details_second chunk.xl

[redacted] Estimation_v0.1.xlsx

[redacted] Revised _Estimation_v0.1.xlsx

Deployment Diagram.png

estimation_ [redacted] xlsx

explaination for presentation.docx

[redacted] xlsx

Invoice File.docx

[redacted] Modernization V1.1 - For review.pptx

[redacted] Modernization V1.1.pdf

[redacted] Modernization V1.pdf

[redacted] Modernization V2.1.pptx

[redacted] Modernization V2.3.zip

[redacted] Modernization V2.4 - PDF.zip

[redacted] DeploymentDiagram.pptx

(A)

You must be very proficient with Wordpress MU 3+, PHP, MySql, CSS...

This could turn into an ongoing project...

You must have sk droid Develope

[redacted]

We have ops, Please place y

Rega [redacted] "We have a che

[redacted] nto a [redacted] k we want to tr

The k is located in: https://github.com/ [redacted]

The

n be developed open-source on Gi

The w valent) to test the

FactorDaily. "Obviously, if people put documents like this in a public GitHub repository, then something has failed — and that failure needs to be looked at."

(B)

Impacts of Exploited Vulnerability

- Economic Exploitation
- Information Theft
- Intrusion on personal privacy
- Social engineering
- System penetration/attack
- Information Bribery
- Sale of personal information
- System sabotage
- Unauthorized system access
- CIA Loss

Research Aim:

Given Crowdsourced Software Development:

- Identify
  - stages involved in confidentiality loss
  - critical attributes of a task
  - type of critical data

- Develop techniques to analyze tasks

INPUT

Task Title
Task Description

OUTPUT

A list of sensitive terms

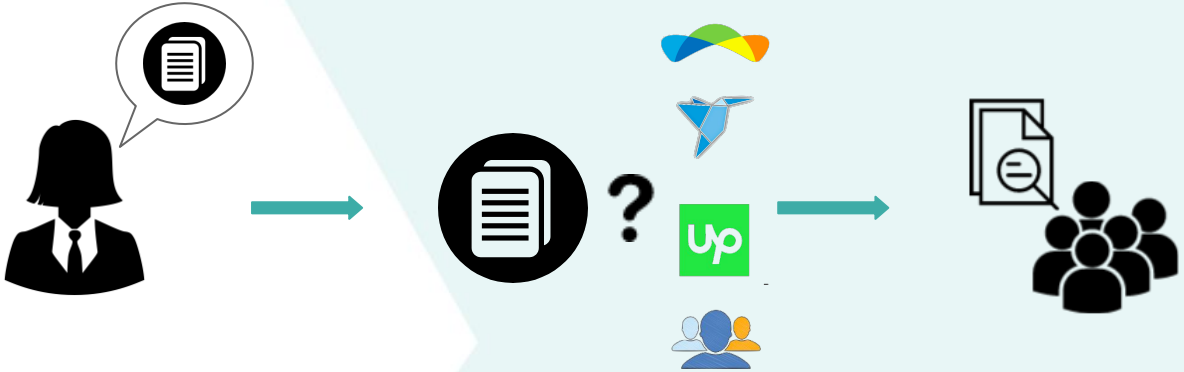
Outline

- Research Motivation
- Research Aim
- Crowdsourcing at the level of organizations
- Survey to understand confidentiality
- Unboxing a typical crowdsourcing task
- Understanding conversations between workers and task posters
- Protecting confidentiality loss in crowdsourcing
- Conclusion

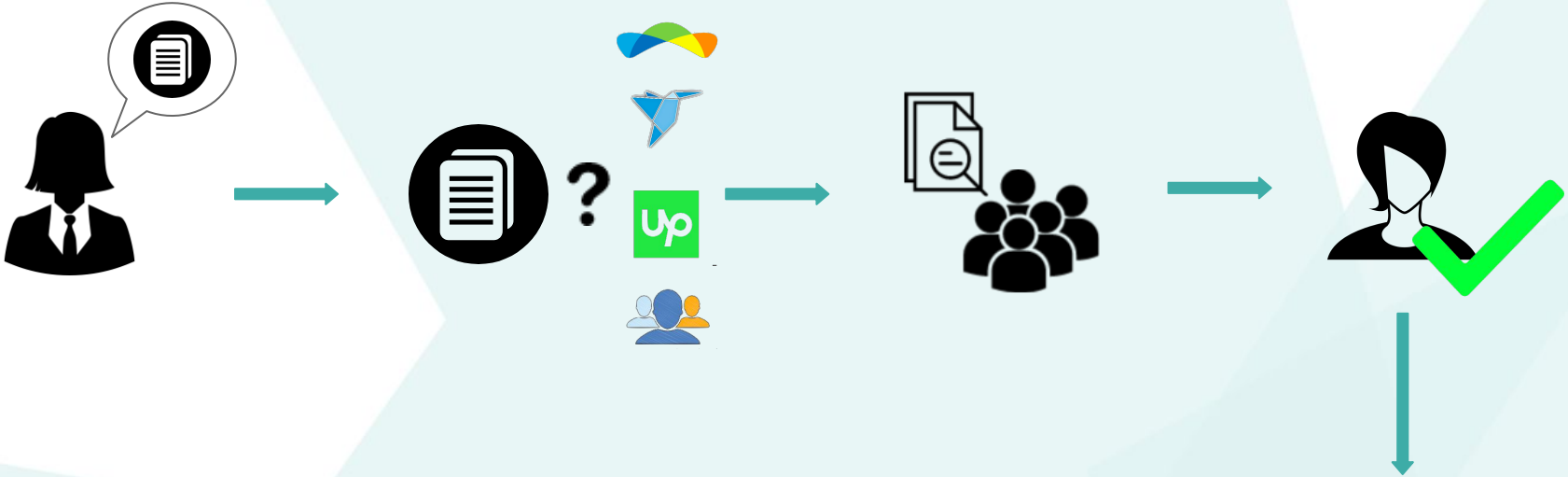
The Crowdsourcing Cycle

The Crowdsourcing Cycle

The Crowdsourcing Cycle



The Crowdsourcing Cycle



The Crowdsourcing Cycle

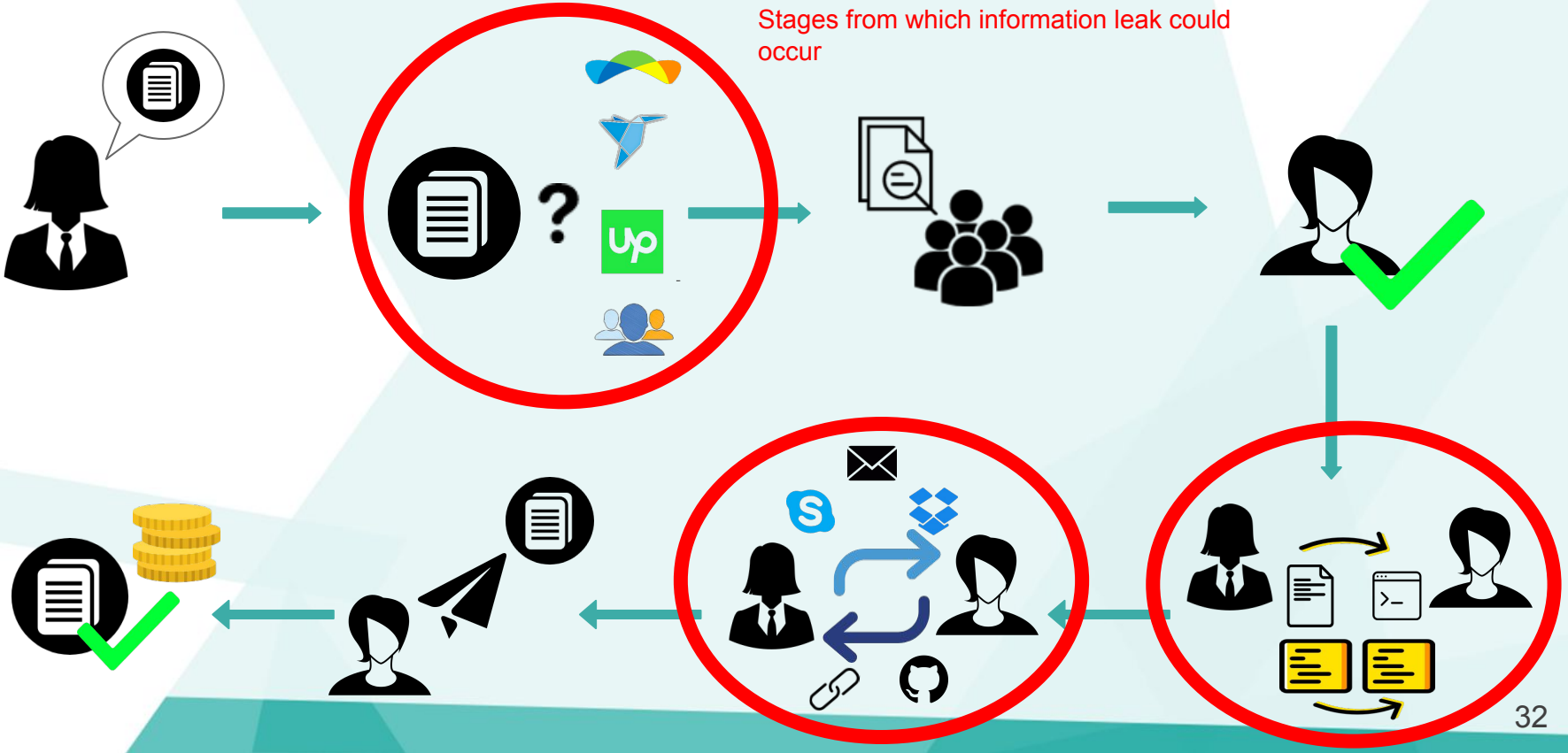
The Crowdsourcing Cycle

The Crowdsourcing Cycle

The Crowdsourcing Cycle

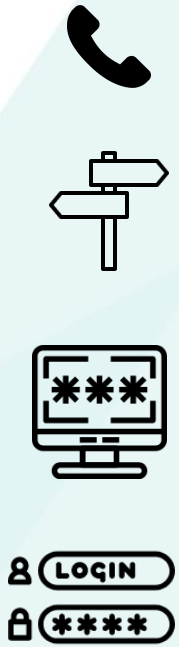
The Crowdsourcing Cycle

Stages from which information leak could occur



Survey to validate some assumptions

- PII details are sensitive
- Sharing a database publicly may compromise with confidentiality
- Sensitive information in code
- Sensitive components in the comments



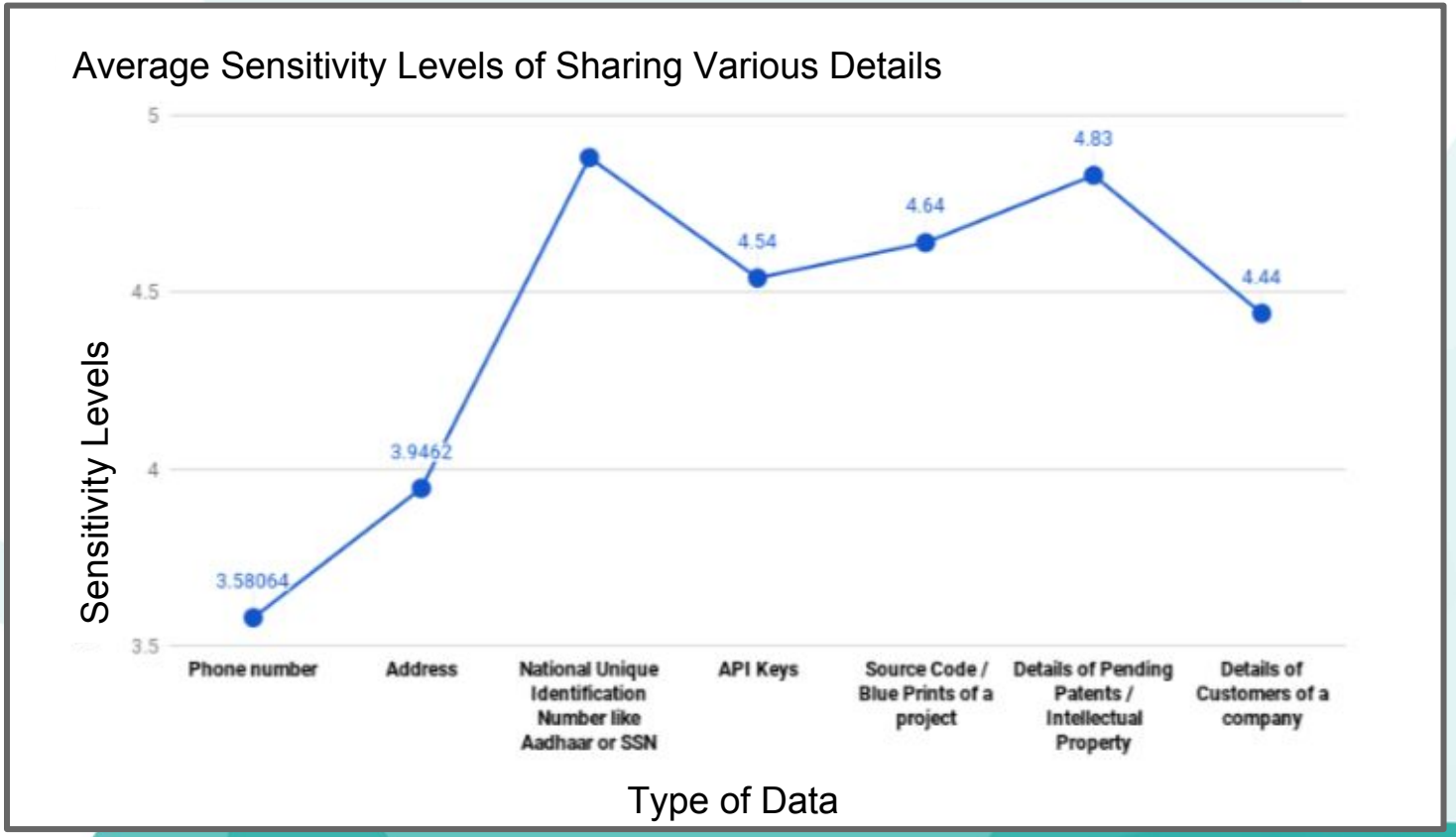
Survey to validate some assumptions

- Wireframes and UX examples may give out information about the company posting the job: logo
- Market of the end result of the task can be inferred given a task
- Domain of the task can be identified from the task, given the type details it is asking for

Survey to understand Confidentiality

Suppose you are a project manager in a multinational organization and you need to get some job done by an external freelancer. These jobs require sharing some details like sample code, documents, UX wireframes, etc. and some information might be sensitive in these resources. You need to answer the following questions considering the above scenario.

Likert scale based Questions



Some observations

- 6.4% of the respondents believed that sharing a database publicly is not a sensitive activity
- 2.2% people think that sharing API Keys is okay

Situational Questions

Sample Task Question

This is a sample task posted by a Task Poster, according to you, what all sensitive data does it contain?

I am using a online platform called tomakeanapp.com, there is a module which you can upload a picture but I HAVE TO DO IT one by one.

I need a imacro or anyscript that could upload a batch of photo from a folder. No other requirement. That's it. It is dead simple and first come first serve. Please apply

If you want to know what i am saying.

Go to testmyapp.com
Username: trialappdemo
Password: password134

>Click my project> Click "Edit", then you will see "TraiApp" in YOUR ACTIVITY

Then click "Edit" and you will see the upload screen I talked above

Sample Task Question

This is a sample task posted by a Task Poster, according to you, what all sensitive data does it contain?

I am using a online platform called tomakeanapp.com, there is a module which you can upload a picture but I HAVE TO DO IT one by one.

I need a imacro or anyscript that could upload a batch of photo from a folder. No other requirement. That's it. It is dead simple and first come first serve. Please apply

If you want to know what i am saying.

Go to testmyapp.com
Username: trialappdemo
Password: password134

>Click my project> Click "Edit", then you will see "TraiApp" in YOUR ACTIVITY

Then click "Edit" and you will see the upload screen I talked above

Sample Code Question

Following is a sample code snippet, point out the parts in the code that you think are sensitive:

```
import pynder

# Replace the given Facebook username and password with yours to get your accees token
# My username -> jakemichael@gmail.com
# My password -> ThIsIsMyPass123
fb_username = ''
fb_password = ''
token =  get_access_token(fb_username, fb_password)

#put in your FB ID here
fb_id = '7003251371123'

# Replace with apli_key
apli_key = 'qpefsey31cs3c24dtahd7sbfs5fshs5w3jESFVB6527ARVJ31LJsglsjsajKbsgdbh426bFvsh'

# To process the paypal pmayment for further authentication to get access to unlimited API calls use my
# Account Number and IPin for payment at www.paypal.com
# Account Number: GH1728K
# IPin: angela52

session=pynder.Session(facebook_token=fb_auth_token,facebook_id=fb_id)


a = 0
for user in session.nearby_users():
    print a
    a += 1
    try:
```

Sample Code Question

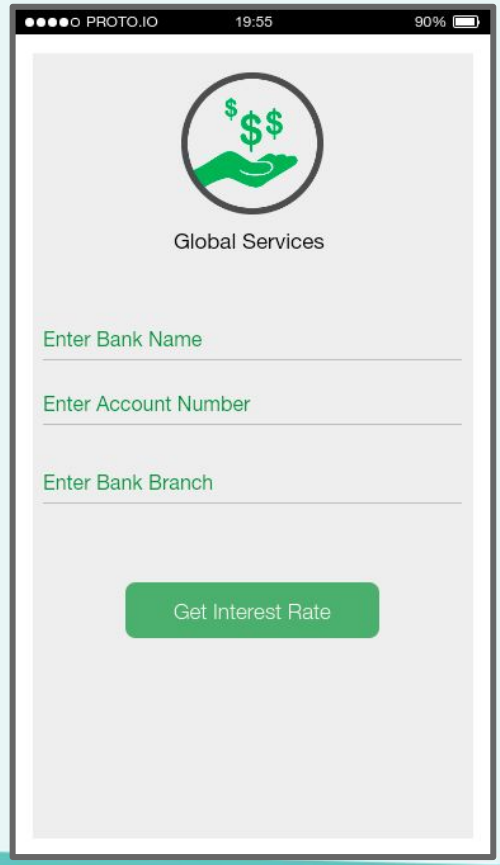
Following is a sample code snippet, point out the parts in the code that you think are sensitive:

```
import pynder

# Replace the given facebook username and password with yours to get your accees token
# My username -> jakemichael@gmail.com
# My password -> ThIsIsMyPass123
fb_username = ''
fb_password = ''
token =  get_access_token(fb_username, fb_password)

#put in your FB ID here
fb_id = '7003251371123'

# Replace with apli key
apli_key = 'qpefsey31cs3c24dtahd7sbfs5fshs5w3jESFVB6527ARVJ31LJsglsjsajKbsgdbh426bFvsh'

# To process the paypal pmayment for further authentication to get access to unlimited API calls use my
# Account Number and IPin for payment at www.paypal.com
# Account Number: GH1728K
# IPin: angela52

session=pynder.Session(facebook_token=fb_auth_token,facebook_id=fb_id)


a = 0
for user in session.nearby_users():
    print a
    a += 1
    try:
```

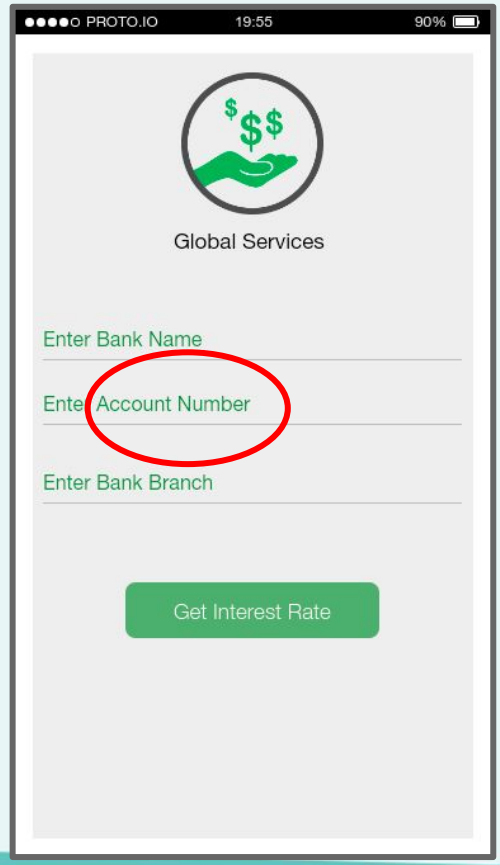
A wireframe of an App has been attached, what all can you infer from it?

A wireframe of an App has been attached, what all can you infer from it?

What can you infer about the type of industry that this app is for?



What can you infer about the type of industry that this app is for?



Other Confidential Entities

1. Master Database of a company

2. Personal Information:
   a. Name
   b. Email ID
   c. SSN / Unique Identification Number
   d. Address
   e. Date of Birth
   f. Employee ID
   g. Health related information

3. Context for the given task

4. Company related details:
   a. Details of the clients of a company
   b. Details of a company
   c. Details of a company's future projects
   d. Any sort of financial data
   e. Proprietary Data

Other Confidential Entities

7. Credentials of any sort: Usernames and Passwords
8. Bank Account Number / Credit or Debit Card Details
9. API Keys
10. Server Details
11. Google Play keys / App Secrets / OAuth Tokens
12. Named Entities
13. Variable names / Class Names
14. Test Inputs
15. Sensitive Datasets where the users can be uniquely identified
16. Images and Diagrams with confidential information

Dataset of 65,000 tasks

- 4 unique Task Categories:
  - Data Science & Analytics
  - IT & Networking
  - Web, Mobile & Software Dev
  - Writing

- 22 Unique Task Subcategories
  - Game Development
  - Data Visualization
  - Network and System Administration
  - Machine Learning
  - ...

Analysis of attributes of a task

- ~ 20 fields per task
  - Task ID
  - Task Title
  - Task Description
  - Task Category
  - Task Sub-Category
  - Date of creation of Task
  - Task Type
  - Task Workload
  - Task Duration
  - Task Budget
  - Preferred Location of the worker
  - Preferred range of feedback score
  - Level of English Skills required
  - Payment range
  - Range of working hours per day
  - Requirement of Cover Letter
  - Requirement of Portfolio
  - Candidates Registered for the task
  - Skills required for the task

Analysis of Tasks

- Crucial attributes:
  - Task Title
  - Task Description

- Observations:
  - ***23%*** tasks contained links
  - ***2.2%*** tasks had credentials
  - **2.46%** tasks had GitHub links

Outline

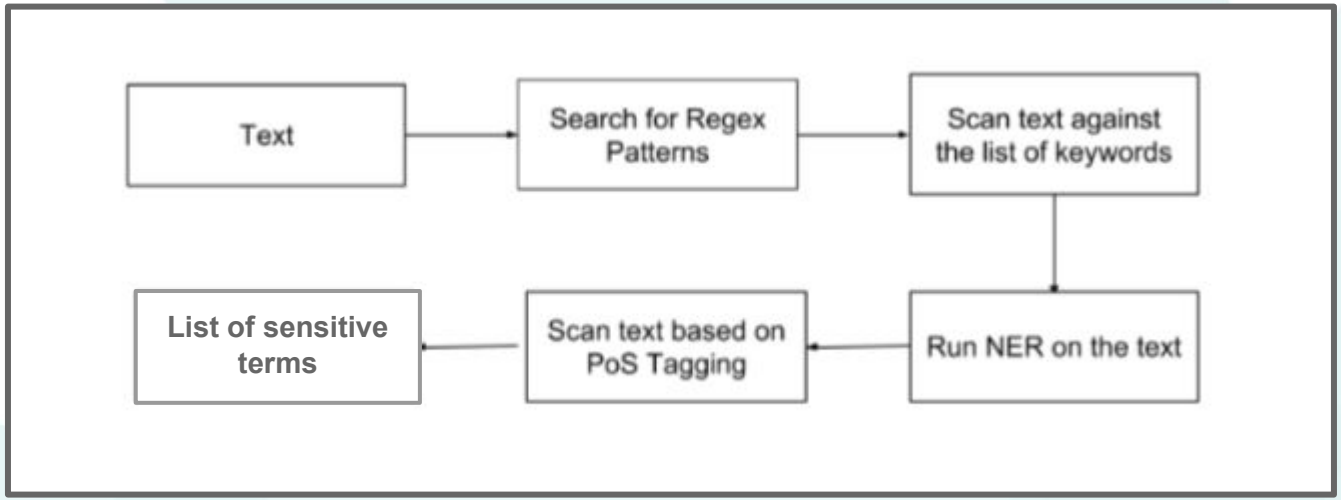
- Research Motivation
- Research Aim
- Crowdsourcing at the level of organizations
- Survey to understand confidentiality
- Unboxing a typical crowdsourcing task
- Understanding conversations between workers and task posters
- Protecting confidentiality loss in crowdsourcing
- Conclusion

Understanding Conversations between Workers and Task Posters

- ~ 214 lines per task (average)
- ~30 files exchanged [min: 0, max: 90]
- 42% tasks involved Skype calls
- 14% involved sharing data over cloud

Analysis of a Task for Sensitivity

Pipeline


Text
Search for Regex Patterns
Scan text against the list of keywords
Run NER on the text
Scan text based on PoS Tagging
List of sensitive terms

Algorithm

Algorithm 1 Analysis of a task

Result: Array of keywords found sensitive in t

Given t and its components tt, td, $r_1, r_2, r_3, \ldots r_n$;

res = [tt, td, $r_1, r_2, r_3, \ldots r_n$];

for *each* ***item*** *in* ***res*** **do**

```
    tokenize(res)
    removeStopWords(res)
    detectRegEx(res)
    detectPasswordsAndAPIKeys(res)
    detectSensitiveKeywords(res)
    detectNER(res)
    add sensitive terms detected to result
end
```

- Task Title ***tt***
- Task Description ***td***
- Task Resources ***r1, r2, r3, ...***

Algorithm 2 Detecting Password and API Key like instances

Result: Return a list of terms identified

```
list = []
for each term do
    if term is not an English Word then
        list.append(term)
    end
    if term is a digit and length(term) > 3 then
        list.append(term)
    end
    if term is alphanumeric then
        list.append(term)
    end
    if term contains special characters then
        list.append(term)
    end
    if term is in CamelCase and has >1 Capital Letters then
        list.append(term)
    end
    Return list
end
```

Testing

Induced sensitivity in the 65,000 tasks dataset

- Curated a list of sensitive terms:
  - Names of companies
  - Email Addresses from the Enron Dataset
  - Dummy API Keys, Passwords, etc
  - List of URLs
  - List of countries and cities
  - Randomly generated usernames
  - ...

Performance Metrics

$$\text{Precision} = \frac{\text{Number of sensitive terms correctly identified}}{\text{Number of sensitive terms identified by the algorithm}}$$

$$\text{Recall} = \frac{\text{Number of sensitive terms correctly identified}}{\text{Number of sensitive terms in the ground truth}}$$

Precision	Recall	F1 Score
0.68	0.82	0.74

Real World Application

- Analyze
  - Tasks
  - Content before putting on cloud
  - Content before sharing

Conclusion

- Narrow down on the crucial stages in the crowdsourcing cycle
- Enumerate critical attributes in a task
- Highlight type of confidential data
- NLP and Rule based algorithm to detect confidentiality loss

Challenges, Limitations and Future Work

- Dataset availability
- Lack of labeled data
- Expand the algorithm
  - Cater to images, databases, etc
- Incorporate Machine Learning techniques for classification
- Sanitization
- More fine grained analysis

Acknowledgement

- Committee Members
- Abhinav and Sakshi, Accenture Labs Bangalore
- Indira, Gurpriya, Arpit, Shubham, Sonu, Divyansh
- Members of Precog family
- Family and friends

References

- http://www.timbroder.com/2015/01/my-2375-amazon-ec2-mistake.html
- https://www.topcoder.com/about-topcoder/customer-stories/
- https://www.merriam-webster.com/dictionary/crowdsourcing
- https://www.figure-eight.com/
- https://www.flaticon.com/authors/freepik

Thanks!

simran13104@iiitd.ac.in

@simran_s21