

## Project Proposal

### Title: **Kayak Alerts for www (1. Proposal)**

This project is motivated by the problem people have to search through large amounts of unstructured data. For instance, if somebody is searching on Craigslist for a set of living room chairs (hence at least 4 or 6 chairs), the user will need to search through more than 100 listings, and spending large amount of time and effort to find suitable listings.

furniture: [by-owner](#) | [by-dealer](#) | both

search for:  in:  ☐ title only ☒ entire post

price:   ☒ has image

sort by [most recent](#) [best match](#) [low price](#) [high price](#)

Found: 1000 Displaying: 1 - 100 [Next >>](#)

[ 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 ]

Sep 23 - [Mid Century Iron Saucer Chair Reproduction](#) - \$45 (lafayette / orinda / moraga) [owner](#) [pic](#) [img](#)

Sep 23 - [★★★7PC Cappuccino Dining Set on Promotion](#) - \$549 [dealer](#) [pic](#) [img](#)

Sep 23 - [FURNITURE SALE!!!! - sofa/bedroom/dining sets - low prices](#) - [dealer](#) [pic](#)

Sep 23 - [Table & 5 Chairs Pine](#) - \$195 (Marin) [owner](#) [pic](#)

Sep 23 - [Patio Dining/Bar Set \(new\) with 4 Bar Chairs and 6 patio chairs](#) - \$375 (fremont / union city / newark) [owner](#) [pic](#)

Sep 23 - [Brand New Solid Teak Chairs - \\$50 each](#) - (oakland downtown) [dealer](#) [img](#)

Sep 23 - [Office Reception Area Sitting Chairs](#) - \$75 (san jose east) [owner](#) [pic](#)

Sep 23 - [EQUINO BARSTOOLS PROVIDE YOU WITH A FUNCTIONAL & STYLISH ACCENT](#) - \$159 (AVAILABLE FOR SHOWING) [dealer](#) [pic](#)

Sep 23 - [The lowest furniture prices in the Bay Area!](#) - (san leandro) [dealer](#) [img](#)

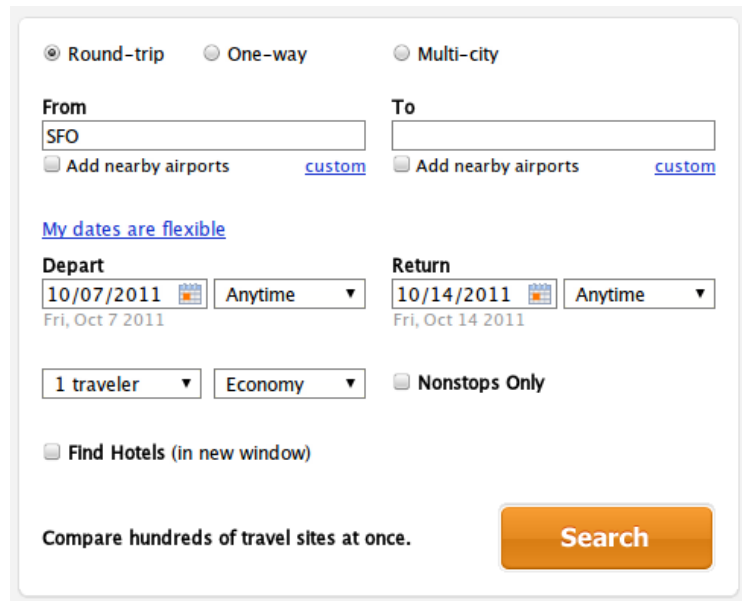
Sep 23 - [Stacking Chairs by VIRCO - Excellent Condition](#) - \$22 (hayward / castro valley) [owner](#) [pic](#)

Sep 23 - [Modern lounge chair and ottoman, recognizable design](#) - \$799 (available for showing) [dealer](#) [pic](#)

Craigslist is just one example out of a large problem sets. Businesses have challenges to automate data retrieval on the web that is ambiguous though essential to their business processes.

The idea of this project is to automate this process by combining the power of human intelligence that Crowdsourcing platforms provide and the ease of automation that web-scrappers provide. In particular, my plan is to develop a system that is as simple as a search for flights on Kayak.com but as powerful as if you personally would have searched through a website. The

way I plan to implement this system is to, first, provide the user with a search interface such as on Kayak.



The image shows a flight search form with the following elements:

- Radio buttons for trip type: ☒ Round-trip, ☐ One-way, ☐ Multi-city.
- From:  To:
- ☐ Add nearby airports [custom](#) ☐ Add nearby airports [custom](#)
- [My dates are flexible](#)
- Depart:    Return:
- Fri, Oct 7 2011 Fri, Oct 14 2011
- ☐ Nonstops Only
- ☐ Find Hotels (in new window)
- Compare hundreds of travel sites at once.
- 

Second, crowdsourcing workers read the query of the user and start searching on the website (around 10-15 searches). Third, a web-scraper learns in the meantime how the structure of the website looks like. Fourth, crowdsourcing workers and the web-scraper search simultaneously through the website to check if the web-scraper has learned the website's structure correctly (around 10-15 searches). Fifth, the web-scraper runs alone with full speed, automating the search effectively.

I believe this project has a significant social component not in the classic 'social/social media' sense, but in the sense that information that are not purely factual are always ambiguous to a certain extend. Ambiguity in that sense, means it requires subjective interpretation which is social process that is depended on culture, gender, economic conditions and other contextual relevant factors.

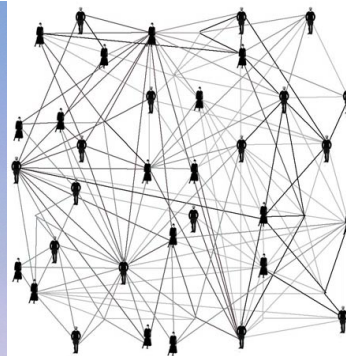
Title: **Malicious Behavior Detection in Crowdsourcing Platforms (2. Proposal)**

One of the biggest challenges that Paypal has been facing over the last couple of years is fraud. More specifically, criminals who would automatically sign up large amounts of paypal accounts, connect these accounts with stolen credit cards, process payments from the credit cards to the paypal accounts and then route this money step by step to their own personal accounts abroad. This problem costs people millions of dollar each year and it is not only paypal who faces this type of malicious behavior.

The solution they came up with is based on a simple observation: Humans have a hard time to behave truly randomly. In particular, the systems and human actions that were used to route money from the malicious accounts to the criminal's account underlies certain patterns. These patterns were used to develop fraud detection algorithms that analysis a broach set of data to identify patterns in the transaction histories of accounts. It turns out that fraud networks look similar to snow flake patterns while usual transaction pattern lack of any clear structures.



Malicious User Pattern



Normal User Pattern

At the same time, current crowdsourcing platforms such as Amazon's Mechanical Turk suffer challenges of large amounts of (intentionally) incorrect or nonsense answers. The reason why this is the case is diverse but one of the reasons can be found in the motivation of workers to maximize their earnings.

The question this projects tries to explore is the idea whether social graph analysis techniques

and fraud detection algorithms can be used in crowdsourcing platforms to identify users who intentionally submit incorrect answers. The problem space is of major importance since malicious worker behavior corrupts the results of projects that businesses and academics have streamlined into crowdsourcing platforms.

The way I intend to approach this project is to study social graph analysis techniques in depth, consult experts in fraud detection algorithms and develop a prototypical system that applies these techniques on the answers (and users) of a project that was submitted to a crowdsourcing platform.

This project has an interesting social component in the sense as it explores how techniques from social graph analysis can be applied in a broader sense and can create value beyond the boarder of social media.