An Evaluation of How Small User Interface Changes Can Improve Scientists' Analytic Strategies

ABSTRACT

Subtle changes in analysis system interfaces can be used purposely to alter users' analytic behaviors. In a controlled study subjects completed three analyses at one-week intervals using an analysis support system. Control subjects used one interface in all sessions. Test subjects used modified versions in the last two sessions: a first set of changes aimed at increasing subjects' use of the system and their consideration of alternative hypotheses; a second set of changes aimed at increasing the amount of evidence collected. Results show that in the second session test subjects used the interface 39% more and switched between hypotheses 19% more than in the first session. They then collected 26% more evidence in the third than in the second session. These increases differ significantly ($p < 0.05$) from near constant control rates. We hypothesize that this approach can be used in many real applications to guide analysts unobtrusively towards improved analytic strategies.

Author Keywords

Visual Analytics, Analytic Biases, Persuasive Technology

ACM Classification Keywords

H.5.2 User Interfaces: Misc

General Terms

Experimentation, Human Factors, Measurement

INTRODUCTION

This paper provides experimental support for the hypothesis that we can use subtle changes in the interfaces of visual analysis systems to influence users' analytic behavior and thus unobtrusively guide them towards improved analytic strategies. We posit that this approach may facilitate the use of visual analytics expertise to correct biases and heuristics documented in the cognitive science community.

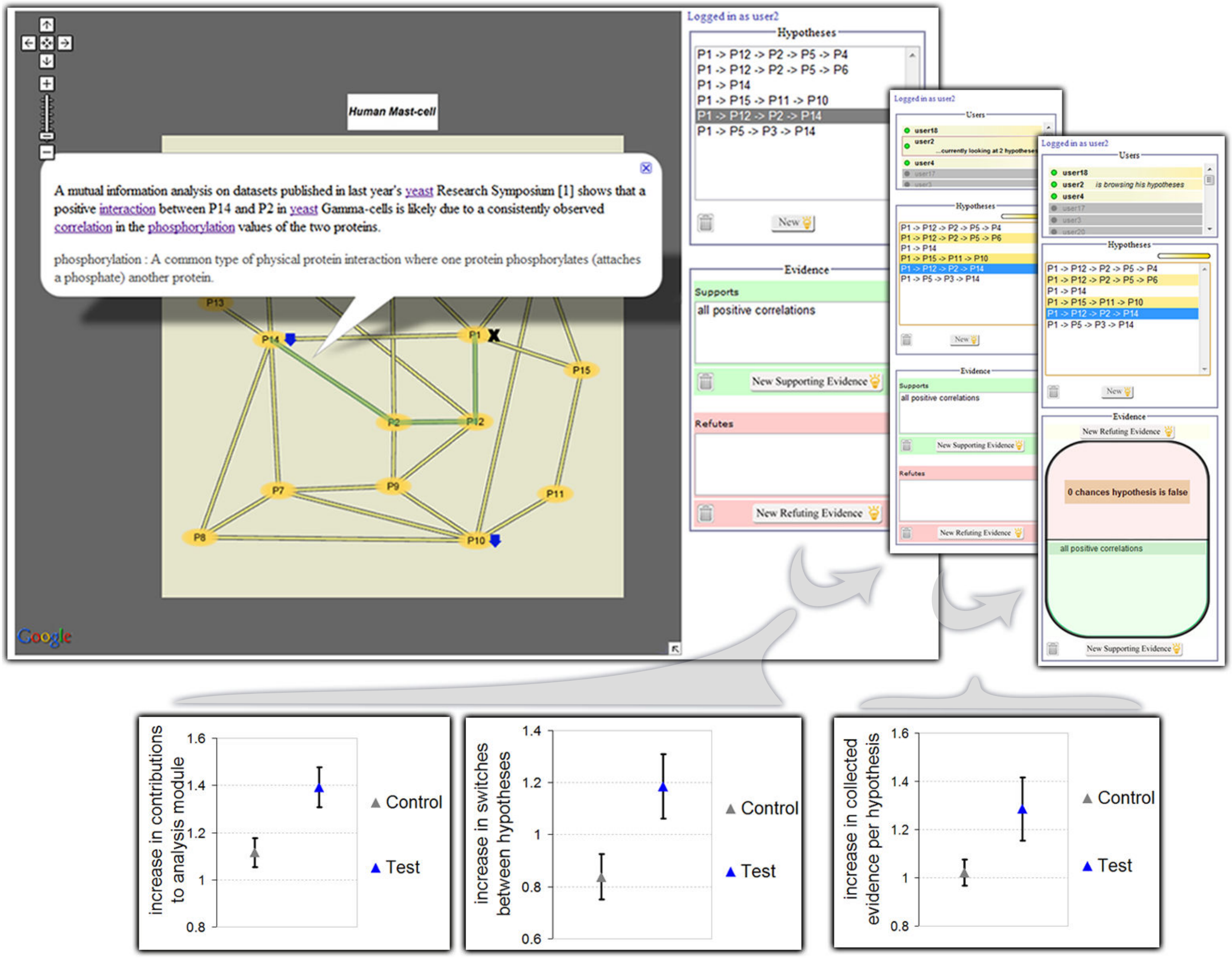
Specifically, we report results from a controlled study in which subjects were asked to complete three analysis sessions using a system consisting of a visualization and an analysis support module. Two sets of non-functional changes were made to the analysis support interface before the second and third sessions. These changes were designed to improve three hypothesized or observed analytic deficiencies: analysts' excessive reliance on memory, inability to consider hypotheses in parallel, and insufficient search for evidence. Our quantitative results show that the interface changes succeeded in alleviating these deficiencies. Compared to a control group, our test subjects used the support module more, they switched among hypotheses more often, and they collected more evidence per hypothesis. Our data not merely show that changes in interfaces translate into different user behavior, but demonstrate that we can leverage interface design and cognitive principles in controlled ways to overcome known analytic deficiencies.

Our work was motivated by extensive cognitive science research showing that human thinking is subject to heuristics and biases that often lead to suboptimal decision making [19]. Recent visual analytics efforts [35] suggest that visualization and interfaces can offer support against such cognitive biases and heuristics, possibly by leveraging the expertise of the cognitive science and intelligence communities [18]. However, to the best of our knowledge, few concrete attempts have used visual analytics techniques to align *descriptive analysis* (i.e., what people actually do to derive a solution) to *normative analysis* (i.e., rational strategies of deriving the best solution). Here, we evaluate a potential solution inspired by previous work in the fields of behavioral economics and human-computer interaction (HCI): *libertarian paternalism* [32, 34] and *persuasive technology* [16] are similar concepts that advocate designing choice layouts and computer interfaces so that they *nudge* users towards decisions that are in their best interest.

Contributions: We hypothesize that subtle changes in visualization interfaces can be used in controlled ways to guide users towards more normative analysis and provide quantitative evidence that supports this hypothesis. We also present qualitative observations on analytic strategies, biases, and heuristics that our subjects used in their tasks.

Roadmap: Next, a related work section summarizes existing research that we build on and extend. We then describe a user-study that validates our hypothesis and summarize its results. We end with a discussion and concluding remarks.

CHI 2012, May 5–10, 2012, Austin, TX, USA.


Human Mast-cell
Logged in as user2
Hypotheses
Users
Evidence
Supports
Refutes
New
New Supporting Evidence
New Refuting Evidence
P13
P14
P1
P15
P2
P12
P7
P9
P11
P8
P10
user18
user2
...currently looking at 2 hypotheses
is browsing his hypotheses
user4
user17
user3
user20
0 chances hypothesis is false
Google
increase in contributions to analysis module
increase in switches between hypotheses
increase in collected evidence per hypothesis
1.6
1.4
1.2
1
0.8
0.6
Control
Test

A mutual information analysis on datasets published in last year's yeast Research Symposium [1] shows that a positive interaction between P14 and P2 in yeast Gamma-cells is likely due to a consistently observed correlation in the phosphorylation values of the two proteins.

phosphorylation : A common type of physical protein interaction where one protein phosphorylates (attaches a phosphate) another protein.

P1 -> P12 -> P2 -> P5 -> P4
P1 -> P12 -> P2 -> P5 -> P6
P1 -> P14
P1 -> P15 -> P11 -> P10
P1 -> P12 -> P2 -> P14
P1 -> P5 -> P3 -> P14

all positive correlations

Figure 1. By making subtle, non-functional changes in the interface of an analysis support module (top) we generated statistically significant changes in users' analytic behavior in a visual problem-solving task. A first set of changes nudged subjects to increase their use of the analysis module by 39% **(lower left,** $p = 0.02$**) in an attempt to support our subjects' working memory. It also caused them to switch among hypotheses** 19% **more often (lower center,** $p = 0.03$**), indicating more consideration of alternative hypotheses. A second set of changes then led subjects to gather** 26% **more evidence per hypothesis (lower right,** $p = 0.01$**). These three increases compare to smaller or negative variations in a control group** $(+15\%, -17\%, -2\%)$**.**

RELATED WORK

This section shows how our work relates to and is motivated by existing research.

Guiding user's choices

Thaler and Sunstein's work [34, 32] in the field of behavioral economics popularized the term *choice architecture* — how a set of choices is presented to a consumer — and the concept of *libertarian-paternalism* — designing choice architectures that "nudge" consumers towards decisions that are in their own interest (paternalistic) while unrestricting choice (libertarian). A similar concept was proposed in the HCI domain by Fogg [16] who defines *persuasive technology* as "interactive information technology designed for changing users' attitudes or behavior". We build on these previous approaches and demonstrate empirically how the nudge paradigm can further the visual analytics agenda.

Sunstein and Thaler as well as Fogg motivate their approaches with two arguments, which they support with experimental evidence. First, any choice architecture or computer interface necessarily influences decision-making behavior, whether intentionally or not. Second, as already shown, research indicates that people's choices and behaviors are not always aligned with their goals. From a visual analytics perspective, this means that even if an analyst's objective is to select the optimal course of action based on available data, cognitive biases and heuristics can steer him towards suboptimal results. Finally, Thaler, Sunstein and Fogg, as well as subsequent research articles, defend the ethicality of the nudging approach and impose ethical design constraints (e.g. avoidance of coercion or deception, ease of avoiding paternalist choices) [25].

Both approaches have inspired scientific results that validated their feasibility. Thaler and Benartzi [33] use an array of cognitive effects such as *mental discounting* or *default options* to persuade employees of a company to increase their contributions to their retirement plans. In the technological realm, the enhanced speedometer [23] changes its appearance based on the current speed limit (when known), encouraging users to stay within speed limits, while the smart sink [2] augments a normal sink with visual cues that make energy consumption apparent. This works has provided inspiring design models for the analysis nudges presented here.

Analytic biases and heuristics

Our work is motivated by extensive cognitive science research demonstrating that people are prone to a range of analysis biases and heuristics that can lead to analysis errors [19]. A specific manifestation of such effects occurs in the context of hypothesis-driven analysis. For example, *satisficing* [29] limits analysis to a hypothesis that is good enough, a *confirmation bias* conditions us to confirm hypotheses rather than disconfirm them [36], and people often fail to to consider alternative explanations [6]. Many such studies have been conducted with naive subjects, but research shows that biases and heuristics also occur in scientific and clinical settings, with similar error-inducing effects [12, 4, 27]. Several results suggest however, that experts, while not immune to such biases and heuristics, may be better equipped to overcome them [13, 21].

Evidence suggests that users can be helped to overcome biases and heuristics and instead use normative analysis, a change that usually yields improved analytic performance. For instance, subjects conditioned to pursue alternative hypotheses and disconfirming evidence reached solutions to a scientific puzzle more often [12]. Analogies and subsequent unexpected findings lead to consideration of multiple hypotheses and novel findings [13]. Medical students using hypothesis-driven analysis outperformed those using a data-immersion approach [15]. Finally, multiple-attribute utility theory can reduce the *prominence* effect (i.e. basing a decision on one attribute deemed most important) [19]. These results support our goal of guiding users towards normative analysis practices.

Related work in visual analytics

The paper "Illuminating the path" [35] introduced the field of visual analytics (VA) as "the science of analytical reasoning facilitated by interactive visual interfaces". The work presented in this paper extends the VA research agenda, concentrating on designing interfaces and visualizations that support the aggregation of data insights into cohesive scientific theories.

Our work is tangential to VA results on understanding the sense-making process [5, 26] and draws on position papers that argue for leveraging the expertise of cognitive science and intelligence communities [18, 31]. In our evaluation we use a system inspired by a range of efforts to design analysis support interfaces that let users store, annotate, and browse analysis artifacts such as hypotheses or evidence [7, 37, 17, 38, 14]. However, to the best of our knowledge, few concrete attempts have used visual analytics techniques to bridge the gap between descriptive analysis and normative analysis. Our work complements current research by using a visual analytics methodology to create a link between observed analytic deficiencies and corrected behavior.

Perhaps closest to our work is that of Savikhin et al. [28] demonstrating experimentally that a targeted visual representation can induce normatively correct decisions in an otherwise biased economic choice task. We extend this result by linking it to the more general nudging approach proposed by Sunstein, Thaler and Fogg, by exploiting interface design in general, and by providing an experimental validation for a high-level analytic task.

METHODS

We conducted a controlled user study to test the hypothesis that small changes in a visualization system's interface can be used to produce targeted modifications in users' analytic workflows. This section presents the design of this study. We start with an overview description of the methodology used and continue with an in-depth presentation of each aspect of the study.

Study overview

Subjects completed an analysis task inspired by a real scientific problem using a visualization and an analysis-support interface (Figure 1, top). Each subject performed three such analysis sessions at one-week intervals. Each session lasted roughly one hour.

Thirty-six subjects, mostly undergraduate and graduate students, were divided into two groups: 21 test and 15 control subjects. The control group solved all three tasks using the same analysis-support interface. Conversely, test-group subjects were given slightly different versions of the analysis-support interface in each session. Specifically, two sets of interface nudges were added to the analysis system before the second and third sessions. We hypothesized that, while changes between sessions would be observed in both groups due to task-learning effects, the test group would exhibit additional effects due to the interface nudges.

The analysis task was inspired by the proteomic domain: finding causal paths in protein interaction networks to explain the interdependency of pairs of proteins that are not directly connected. None of the subjects was familiar with the task or background material beforehand and all received a 20-minute tutorial at the beginning of the study.

Our test system was instrumented to log users' interactions automatically. Subjects were also asked to distill their analysis in a written questionnaire at the end of each of the three analysis sessions. We analyzed the datasets both quantitatively, to look for support for our nudging hypothesis, and qualitatively, to gain insight into how subjects approached their task.

Task description

Subjects were asked to solve three artificially constructed analysis tasks inspired by workflows of proteomic researchers studying protein signaling pathways.

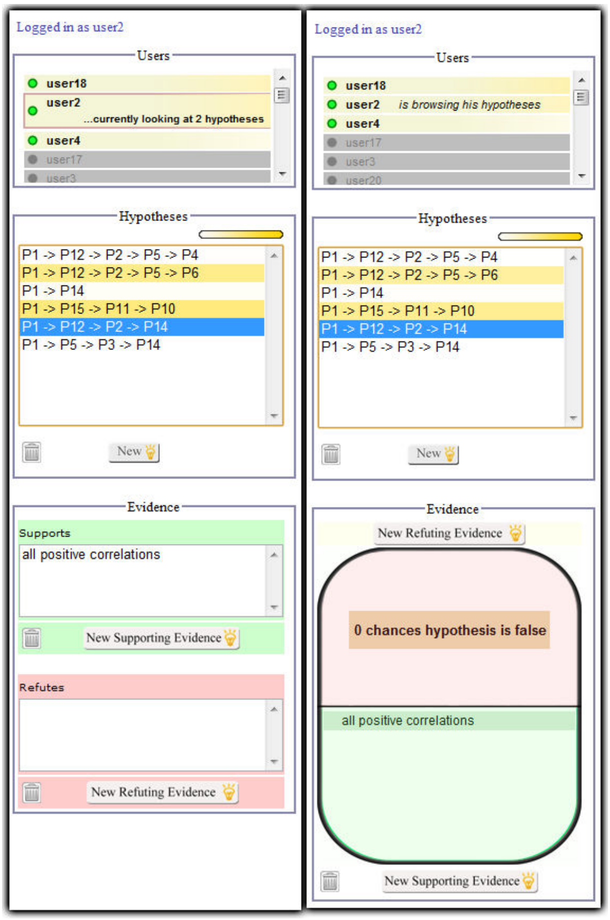
Proteins are functional molecules within cells that interact with one another to form complex causal pathways that determine the response of cells to events. Such protein interactions are the object of intense scientific research because understanding these cellular pathways would let researchers devise efficient drugs to influence a cell's behavior without causing unwanted side effects. Proteomicists often use visualizations of interaction networks to understand changes in protein activation patterns measured in experiments. A distinct class of experiments is *knockout experiments*: here researchers deactivate particular proteins and compare protein activation levels before and after the removal.

Our subjects were given network visualizations that were said to depict protein interactions documented in recent publications. Figure 1 shows one of three distinct networks that subjects were asked to analyze. The networks were manually created and laid out. The familiar Google Maps user interface was used to display the network images and offer basic interaction. Clicking on nodes or edges opened information bubbles referring to these particular elements. Interactions were described by fictional, brief paper abstracts detailing the particulars of each interaction and the context in which it was discovered.

Subjects were told that a knockout experiment had been performed on a specific type of cell. They were informed that a protein was removed from the cell and that researchers subsequently observed changes (positive or negative) in the levels of several proteins. These changes were marked on the network with arrows. Finally, subjects were asked to use the available information to determine network paths likely to have produced those changes and to rank them in order of plausibility. This network task represents a visual, complex, and open-ended implementation of causal reasoning tasks that have been typical choices of cognitive studies [36].

Our networks used proteomic terminology but introduced fictitious proteins, interactions and interaction mechanisms. Thus, the probability of a regulation chain was determined by the logical consistency of the evidence presented. The key rules that subjects were expected to extract from the evidence and use in their analysis were: the probability of a depicted interaction is lower if it was documented in species and cells other than those investigated in the knockout experiment; a correlation between two proteins should be treated as an edge with uncertain directionality; interactions could describe direct or inverse regulation mechanisms; and the edges sequence in a solution path should justify the sign of the observed change. These assumptions, along with a general description of protein signaling, were illustrated in a 20-minute tutorial (text and video) and were clarified on request. Moreover, essential terms were highlighted in all evidence text and in-situ explanations were displayed upon mouse clicks (Figure 1).

The order in which the three networks were presented to users was alternated to minimize the chance of network differences influencing the global result. Thus, in the test group, six subjects solved the networks in order 1, 2, 3, seven subjects solved them as 2, 3, 1, and the remaining six solved them as 3, 2, 1. A similar division was used for the control group.



P1 -> P12 -> P2 -> P5 -> P4
P1 -> P12 -> P2 -> P5 -> P6
P1 -> P14
P1 -> P15 -> P11 -> P10
P1 -> P12 -> P2 -> P14
P1 -> P5 -> P3 -> P14

Figure 2. The two modified analysis interfaces include three evaluated nudges: a box lists online users actively interacting with the analysis module (left), a color gradient (white to gold) shows recently analyzed hypotheses (left), a redesigned, larger evidence box asks users to commit to the implications of a hypotheses lacking associated evidence (right).

Analysis interface and evaluated nudges

In addition to the protein network viewer, an analysis support interface augmented the experimental environment (Figure 1). As noted in Section 3.1, control subjects used the same base analysis interface in all three sessions. Test subjects started with an identical interface but then used upgraded versions in the second and third sessions. These versions, obtained by incrementally including two sets of evaluated nudges in the base interface, are shown in Figure 2.

The base analysis module contained three lists in which users could store their hypotheses, confirming, and disconfirming evidence. Hypotheses were entered into the system as noncyclical network paths by clicking on sequences of connected nodes. Evidence was inserted into the confirming or disconfirming category by typing free text in a pop-up box.

Selecting existing hypotheses would highlight their corresponding paths on the visualization and display their associated evidence, thus allowing subjects to revisit and compare hypotheses. Subjects were familiarized with these features in the tutorial video at the beginning of the study.

Three nudges were designed to alleviate three analytic deficiencies. First, we assumed that subjects would rely on their working memory rather than use the analysis system. Second, based on cognitive science studies, we assumed that subjects would have trouble considering multiple hypotheses in parallel. Third, we hypothesized that subjects would gather mostly confirming evidence for their hypotheses, and ignore the aspect of disconfirming evidence. The results of our initial session one runs caused us to adjust our last assumption: subjects were gathering approximately equal amounts of confirming and disconfirming evidence but in overall small amounts. We refined the design of our last nudge to target this issue better.

As already noted, the **first evaluated nudge** (Figure 2, left) aimed to increase users' reliance on the analysis module. Our design rested on the assumption that if subjects knew other users were actively interacting with the module, they would do so as well. To test this assumption, a section listing online users was added to the base analysis module. As users interacted with the module, this was reflected in a publicly visible status message (e.g., "user is browsing his hypotheses", "user has entered new evidence"). Fake user-bots were added to ensure that a nudge factor was present at all times. This design was inspired by research on conformity effects and motivational factors for online contributions. Specifically, humans change their behavior to match that of others [9, 3, 8] to gain social approval [10], or because they derive utility information from observing what others do [30, 20]. In addition, visibility encourages users of social networks to increase their online contributions [1, 24, 24].

The **second nudge** was designed to encourage users to compare and contrast hypotheses in parallel rather than perform a sequential search in hypothesis space. Initially we planned to evaluate this nudge by itself but ultimately merged it with the first one so as to make the length of the study manageable. The design involved assigning each hypothesis a recency score that decayed over time but increased with any interactions targeting the hypothesis (e.g. selection, adding evidence). Recently active hypotheses were highlighted in the hypotheses list by using a color gradient based on the recency score. Finally, thresholding the recency score allowed us to determine the number of a user's active hypotheses, display this information in the user status (Figure 2, left), and sort users based on how many hypotheses they were investigating. This offered a visual and status reward. While a user could trick the system by quickly switching between hypotheses, this was taken into account in data analysis (see Results section) and we have observed just two intentional instances of it. These first two nudges were integrated into the analysis interface before the second session.

Finally, the **third nudge**, deployed before the last session, aimed to encourage test subjects to gather more evidence for the same number of hypotheses. To that end we modified the evidence collection part of the interface (Figure 2, right). First, the evidence-collection area was made more visually interesting and distinct from the rest of the interface. Second, if no confirming or disconfirming evidence had been entered for a hypothesis, the evidence boxes would read "0 chances that hypothesis is false" or "hypothesis is unlikely". This essentially required subjects to commit to extreme cases — something that people are known to avoid [11]. Third, modification introduced unintentionally while implementing the design was that the evidence boxes in this nudge were larger than in the base interface.

We hypothesize that this nudge could be restricted to disconfirming evidence only, in which case it could potentially alleviate confirmation biases [36]. As noted, our subjects did not exhibit a confirmation bias in the early stages of the study, so that we resorted to testing the more general case of increasing the amount of total evidence.

User pool

Our study included a total of 36 subjects. Of these, 16 were women and 20 men. Six of them were young professionals, 18 were undergraduates, and 12 were graduate students. Twenty-six of the subjects were active in sciences, while 10 were humanities students. None of the subjects had previous experience with proteomic analysis. Thus, all subjects relied solely on the tutorial provided at the beginning of the study.

Subjects were randomly distributed in control (15) and test (21) groups such that the two groups had similar distributions of gender and age (undergraduate, graduate or postgraduate). Subjects were compensated for their participation.

User study limitations

Ease of hypothesis elicitation: A pilot run showed us that free-text specification of hypotheses would have produced considerable variability in what users entered as hypotheses. To be able to compare results across subjects we limited hypotheses to paths of connected proteins. This interaction mode, reinforced by the tutorial video, gave subjects an easy "recipe" for generating hypotheses: any network path was a valid hypothesis.

Lack of motivation: Our study did not involve monetary incentives to encourage subjects to provide valid solutions. As a result, several subjects appeared not to devote significant effort in searching for clues beyond those immediately noticeable.

Unforeseen problem-solving strategies: A few of our early subjects copied the network on paper and annotated each interaction and protein. This strategy is not scalable to real protein interaction networks and it does not capture the exploratory nature of analysis. To avoid this, we instructed the rest of the subjects not to use such exhaustive analysis strategies.

Task misunderstanding: Instead of constructing short paths that linked the knockout protein to each changing protein, two subjects looked for long paths that linked the knocked-out protein and all arrow proteins (i.e proteins with changed levels) together. We retained these results because the subjects used this interpretation consistently in all three sessions.

Variation in analysis times: We urged users to spend approximately 60 minutes on each session. Several subjects, however, insisted on finishing earlier. Moreover, a few datasets showed prolonged intervals of inactivity and several users were observed to take web-browsing or texting breaks. In our analysis we eliminated intervals with no activity and normalized all measurements by the time spent on the task.

Small number of subjects: Our sample size (21 test, 15 control subjects) was relatively low for the open-ended tasks our study involved. However, we note that the trends in the data became apparent with as few as six users in each group and changed very little throughout the experiment.

Effect of change not captured: Our study does not capture the amount by which interface changes amplify the saliency of our nudges. It may well be that nudges are less observable and effective if they are introduced into the first system release.

RESULTS

Here we describe quantitative and qualitative results from our user study. All data and analyses are available online [39].

Data preparation and analysis

Thirty-two subjects completed all three sessions while four completed only the first two for a total of $28 \times 3 + 4 \times 2 = 104$ datasets. Four of the subjects, two from each group, solved the tasks on paper using exhaustive annotation of the networks. Three additional users also switched to this approach in the final session. All these data were discarded from the analyses leaving $104 - 4 \times 3 - 3 \times 1 = 89$ datasets from 13 control subjects and 19 test subjects.

We measured and analyzed three quantitative indicators to support our nudging hypothesis. First, we recorded the number of hypotheses and evidence entered into the system as a proxy for the degree to which subjects relied on the interface to trace their analysis. This number was normalized by the time, in minutes, subjects spent on each session. Second, we measured the number of times a subject switched between hypotheses and normalized it by the number of hypotheses, as an indicator of the degree to which hypotheses were analyzed in parallel during analysis. Third, we recorded the number of evidence items collected and divided it by number of hypotheses.

In the case of hypotheses switches, we ignored hypotheses selections lasting less than 5 seconds because we observed that users sometimes cycled rapidly through hypotheses as a method of gauging progress. We also ignored switches occurring in the last part of the analysis while subjects were filling in the answer questionnaire. We found that by default most users did a comparative analysis of hypotheses at the very end. Our nudge however was designed to encourage users constantly to consider alternatives.

In a second phase we also made a qualitative analysis of our subjects' workflows. Our goals were to understand the dominant analytic strategies and behavioral patterns, and to verify the degree to which biases and heuristics were applied.

Quantitative support for nudging hypothesis

The premise of our experiment was that interface nudges would cause test subjects to change their behavior between sessions differently from how control subjects' behavior would evolve naturally as a consequence of learning or boredom. Figures 3-5 demonstrate the validity of our premise by contrasting the relative changes in performance measures between consecutive sessions in both experimental groups. As expected, change was negligible in control subjects (means of all triangles are close to one), but was significant for test subjects when a nudge was present (means of black squares greater than one). However, test group behavior remained constant whenever performance measures were not specifically targeted (e.g. change in contributions between the last two sessions). This suggests that subjects were responding not simply to interface changes but instead to nudges targeting particular performance measures.

Test subjects contributed 39% more hypotheses and evidence items to the analysis module in the second session than in the first. This compares to an increase of only 15% in the control group (Figure 3). A t-test found this difference to be statistically significant ($t(29) = -2.07, p = 0.02$). Contributions remained close to constant between the second and third sessions in both the control and test group (Figure 3). This conforms to the expected behavior since no nudge targeting contributions was added between these sessions.

The difference in switches between hypotheses was an increase of 18% in test subjects versus a decline of 17% in control subjects (Figure 4). The difference was significant, as indicated by the t-test ($t(25) = -1.89, p = 0.03$). The first two nudges were both added before the second session. Thus, we assign either of the observed changes not to any single nudge but to all interface changes made between the first two sessions.

The amount of evidence collected per hypothesis remained fairly constant between *all* consecutive sessions in the control group with a decrease of 2% (Figure 5). Test subjects however, gathered on average 24% more evidence per hypothesis in the third condition than the second. This difference was also found to be statistically significant ($t(38) = -2.28, p = 0.01$).

Qualitative analysis of subjects' workflows

Our subject's logs allowed us to assess their workflows qualitatively to extract common strategies and to determine the extent to which subjects relied on analytic biases and heuristics. The following paragraphs summarize our conclusions.

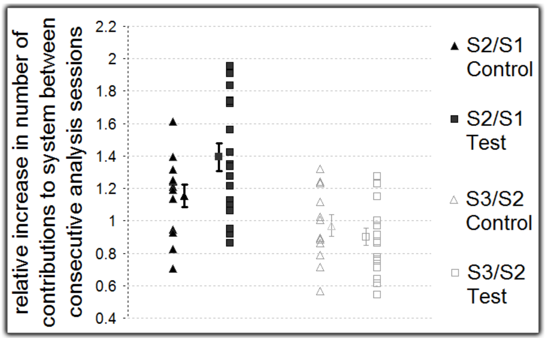


Figure 3. Changes between the first two sessions (black) caused test subjects (squares) to increase the number of hypotheses and evidence items entered into the analysis system by an additional 24% over the control subjects'(triangle) relative increase. The interface changes before the third session had no significant impact on this performance measure (gray).

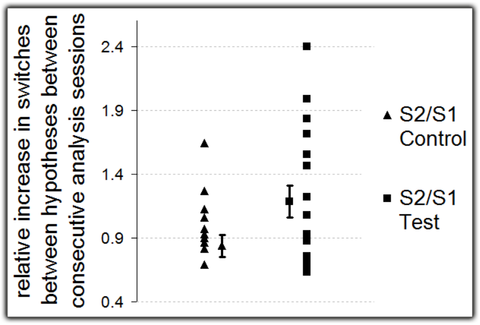


Figure 4. Changes between the first two sessions caused test subjects (squares) to switch between hypotheses an additional 35% more than than the control subjects'(triangle) relative increase.

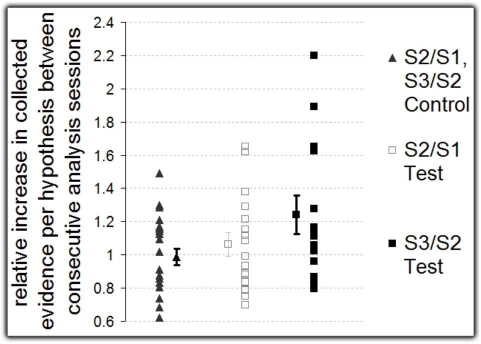


Figure 5. Changes between the last two sessions (black) caused test subjects (squares) to gather 24% more evidence for their hypotheses as opposed to a constant evidence/hypotheses ratio (-2%) between all consecutive control sessions (triangles). Changes in the test group before the second session (gray) produced non-significant changes in evidence collection as compared to the control group.

Observed workflows: More than half our subjects started with an initial exploration of the network. This exploration was not hypothesis driven and typically lasted between three and six minutes. Subjects then moved on to a hypothesis-driven analysis, trying to connect arrow proteins to the knock-out protein (Figure 1). We could discern two strategies for entering hypotheses. Most subjects would pick a candidate path, do a pre-evaluation of its likelihood, enter it into the system if it was plausible, and then follow with a second pass to summarize and document evidence. These users would often revisit hypotheses and compare them. A few subjects added hypotheses without prior exploration and then summarized evidence in a following pass. Generally, they did not reevaluate those hypotheses again until a final pass when they decided on a global likelihood ordering.

Observed biases and heuristics: An interesting finding was that confirmation bias was not dominant. In fact, subjects gathered slightly more disconfirming evidence than confirming evidence. A number of qualitative observations provide further support for this finding. First, several users gathered almost exclusively disconfirming evidence, while others pruned paths that had strong negative evidence. Second, one subject would copy entire sections from the information bubbles and enter it directly as confirming evidence, but would always carefully summarize negative evidence. This suggests that she recognized the higher diagnostic value that the disconfirming evidence would have in her final ranking.

A known heuristic that we found in several datasets was *single-attribute analysis* (i.e. focusing on a single most prominent attribute and using that to rank options). We noticed several cases in which subjects added complicated paths before shorter, more intuitive ones. On closer inspection we found that they had selected a single attribute (e.g. cell type) and were using it to include or discard paths from their analysis.

We also noticed an inability to operate with varying degrees of probability. Several subjects seemed to postpone the consideration of paths involving a complex probability judgment (e.g. multiple interactions with associated uncertainty) and instead concentrate on paths that allowed a binary decision.

Our network setup was well suited to discovering *conjunction fallacies*, which occur when a specific condition is deemed more likely than a general one. In our network task short paths should be more likely candidates for analysis than longer paths. In general, our subjects seemed aware of this principle. In fact most new hypotheses abided by this rule. Additionally, several subjects added the short length of a path as positive evidence. However, we noticed that subjects' analytic strategies tricked them into the conjunction fallacy in a significant number of cases. We observed three main scenarios leading to this.

First, the favored method of expanding a set of hypotheses was to modify an existing one by rerouting part of its path. At the very least, subjects often used interactions that they were already familiar with. Most subjects avoided picking completely new routes, especially in network areas where they had already done some analysis. Such small changes to initially short paths lead subjects to analyze increasingly longer paths. Ultimately, subjects spent considerable time on long paths that were less likely than unexplored shorter options.

Second, subjects occasionally considered longer paths linking together multiple arrow proteins more likely than short paths from the knockout protein to each of those arrow proteins. We hypothesize that users were looking for good unifying stories, a known cognitive tendency. Interestingly, one of the subjects confessed that he was aware of the conjunction fallacy but that the "story was too good" to be irrelevant.

The third reason for multiple instances of conjunction fallacy was tied to the network layouts. The way paths were visually displayed had an impact on which ones were chosen for analysis. Most subjects preferred paths that described fairly continuous visual arches, or that were symmetric with ones they had already looked at. Sharp-angled paths were usually selected last even if they were shorter than already analyzed hypotheses. Another effect observable in several datasets was that symmetrical paths were more often compared to each other than to other hypotheses.

DISCUSSION

Here we discuss the broader impact of our contributions, alternative methodologies, and open questions.

Significance

Some of the findings reported in this paper may seem unsurprising. That interface design can alter analytic workflows is evident, as is the fact that online visibility is correlated with increased online activity [24]. However, our study data shows more than that interface changes translate into different user behavior. Our contribution lies in demonstrating that interface elements can be leveraged in controlled ways to unobtrusively correct users' strategies: our subjects' deficiency in supporting their hypotheses with evidence was observed in the first session and alleviated by a redesigned analysis-support module in the third session. We believe this approach is valuable because it has the potential to correct and improve users' strategies without having to rely on coercive or obtrusive elements such as pop-up messages or help agents.

Design guidelines

Our work was primarily aimed at providing experimental support for the nudge paradigm in the visual analytics domain rather than providing a set of design guidelines. The nudge design space warrants more exhaustive exploration because it can either provide a tool for guiding users towards better analytic strategies or help us understand how our interfaces unintentionally shape users' exploratory and analytic patterns. Our work actualized interesting questions about the degree to which tutorials, ways of entering and storing hypotheses, and even simple design choices such as text-area size and color can influence users' behavior.

A few loose design guidelines, however, can be distilled from our work. First, putting collaborative elements and conformity triggers in analysis systems can nudge users to change their behavior. We hypothesize that artificial *model-analysts*, such as used in our experiment, could nudge users towards conforming to a desired behavior. Second, visual rewards, such as our recency score, will encourage users to consider options in parallel. Third, messages in text areas, perhaps in conjunction with box size, may be used for boxes that should not be left empty. Finally, from our qualitative analysis of our subjects' workflows we hypothesize that ways of automatically suggesting hypotheses may alleviate some of the observed conjunction fallacies and that subjects would benefit from support for multiple attribute analysis. Both such mechanisms would need to be domain specific and are beyond the scope of the present work.

General considerations

The data distributions may suggest that nudges, rather than uniformly targeting all subjects, tend to be particularly effective for a subgroup and less so for the rest. As seen in Figures 3-5, measurements obtained from test subjects appear to form two clusters: one with values similar to those measured in the control group, and one with distinctively higher values. These clusters do not correlate with the order in which networks were presented to users. However, the data gathered as part of this study is insufficient to test this hypothesis.

The analytic biases and heuristics targeted in our study were chosen because they are amply documented in the cognitive science literature. It is likely that one or more of these effects do not appear or are beneficial in some areas or settings. In fact *naturalistic decision making* [22], a distinct research area, models situations (e.g., crisis control, time-sensitive operations) in which heuristics are an efficient analytic strategy. The aim of this study was not to eliminate a specific set of biases and heuristics but to demonstrate that if such effects are identified we can use interface elements to reduce their occurrence.

Our study did not replicate several biases and heuristics documented in the cognitive science literature. Most notably, people are thought to be unable to elicit many hypotheses and to be biased towards gathering predominantly confirming evidence. Conversely, our subjects generated many hypotheses and showed no confirmation bias. We see two possible explanations for this. First, two of our study limitations may be responsible: the ease of generating hypotheses and subjects' lack of motivation led them to pursue multiple hypotheses and not develop attachments to favored ones. An alternative explanation is that people can switch from a normal working mode to an analysis mode in which normative principles are more carefully observed. Research by Dunbar [13] hints at this hypothesis.

This latter possibility supports our choice of analysis task. Shorter and more focused tasks like the ones used in many cognitive experiments can be applied to large numbers of users and provide clean data. However, the extent to which they translate to the exploratory analysis typical of scientific discoveries is far from clear. As noted in the related work section, several studies indicate that there are observable differences between laboratory settings and real scientific or clinical situations.

Similarly, our study might have been more informative had we tested domain experts in their field of research rather than naive users on unfamiliar tasks. It remains uncertain whether domain experts, who generally follow well established workflows, can be nudged as easily as our subjects. Moreover, a high familiarity with an analysis system may also cause expert subjects to overlook new interface nudges.

Unfortunately, domain experts are scarce and the variability in the scientific problems they solve is high. Thus, quantitative studies that faithfully replicate real-life scientific settings are scarce and likely to remain so. Our choice of task and users implements a realistic approximation that provides insight into how to minimize the impact of biases and heuristics in scientific workflows. This endeavor is important because, as remarked at the beginning of the paper, domain experts are not immune from cognitive biases and heuristics and often benefit from normative analysis strategies.

CONCLUSION

We presented results from a quantitative user study demonstrating that controlled changes in the interface of an analysis system can be employed to correct potential deficiencies in users' analytic behavior. Specifically, we manipulated the design of a basic analysis tool over three analysis sessions to produce three changes in our subjects' analysis. First, subjects were nudged to increase their reliance on the analysis-support module that accompanied the visualization. Second, subjects were nudged to analyze hypotheses in parallel rather than sequentially. Third, subjects were nudged to gather more evidence for their hypotheses. The significance of our work is threefold. First, we give an account of how even the simplest design decisions shape users' analytic behavior. Second, we advance visual analytics efforts by introducing and validating an approach that leverages visualization environments to correct analytic biases and heuristics reported in the cognitive science literature. Third, we provide a short overview of analysis workflows, and biases and heuristics that our subjects used on a scientifically inspired analysis task.

REFERENCES

1. Ames, M., and Naaman, M. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, ACM (2007), 980.
2. Arroyo, E., Bonanni, L., and Selker, T. Waterbot: exploring feedback and persuasive techniques at the sink. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, ACM (2005), 639.
3. Asch, S. Studies of independence and conformity: a minority of one against a unanimous majority. *Psychological monographs 70*, 9 (1956), 1–70.
4. Ben-Shakhar, G., Bar-Hillel, M., Bilu, Y., and Shefler, G. Seek and ye shall find: Test results are what you hypothesize they are. *Journal of Behavioral Decision Making 11*, 4 (1998), 235–249.
5. Bodnar, J. Making sense of massive data by hypothesis testing. In *International Conference on Intelligence Analysis* (2005), 2–4.
6. Bruner, J., and Potter, M. Interference in visual recognition. *Science 144*, 3617 (1964), 424–425.
7. Canas, A., Carff, R., Hill, G., Carvalho, M., Arguedas, M., Eskridge, T., Lott, J., and Carvajal, R. Concept maps: Integrating knowledge and information visualization. *Lecture Notes in Computer Science 3426* (2005), 205.
8. Chartrand, T., and Bargh, J. The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology 76*, 6 (1999), 893–910.
9. Cialdini, R., and Goldstein, N. Social influence: Compliance and conformity. *Annual Psychology Review 55* (2004), 591–621.
10. Deutsch, M., and Gerard, H. A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology 51*, 3 (1955), 629–636.
11. DuCharme, W. Response bias explanation of conservative human inference. *Journal of Experimental Psychology 85*, 1 (1970), 66–74.
12. Dunbar, K. Concept discovery in a scientific domain. *Cognitive Science 17*, 3 (1993), 397–434.
13. Dunbar, K. What scientific thinking reveals about the nature of cognition. *Designing for science: Implications from Everyday, Classroom, and Professional Settings* (2001), 115–140.
14. Eccles, R., Kapler, T., Harper, R., and Wright, W. Stories in geotime. *Information Visualization 7*, 1 (2008), 3–17.
15. Elstein, A., and Schwarz, A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *Stroke 33* (2002), 493–6.
16. Fogg, B. Persuasive computers: perspectives and research directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM Press/Addison-Wesley Publishing Co. (1998), 225–232.
17. Gotz, D., Zhou, M., and Aggarwal, V. Interactive visual synthesis of analytic knowledge. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology* (2006), 51–58.

18. Green, T., Ribarsky, W., and Fisher, B. Building and applying a human cognition model for visual analytics. *Information Visualization 8*, 1 (2009), 1–13.

19. Hastie, R., and Dawes, R. Rational choice in an uncertain world. *Journal of the Indian Academy of Applied Psychology* (2003), 107.

20. Kenrick, D., Maner, J., Butner, J., Li, N., Becker, D., and Schaller, M. Dynamical evolutionary psychology: Mapping the domains of the new interactionist paradigm. *Personality and Social Psychology Review 6*, 4 (2002), 347.

21. Klayman, J., and Ha, Y. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review 94*, 2 (1987), 211–228.

22. Klein, G. A recognition-primed decision (rpd) model of rapid decision making. *Decision Making in Action: Models and Methods* (1993), 138–147.

23. Kumar, M., and Kim, T. Dynamic speedometer: dashboard redesign to discourage drivers from speeding. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, ACM (2005), 1576.

24. Nov, O. What motivates Wikipedians? *Communications of the ACM 50*, 11 (2007), 64.

25. Oinas-Kukkonen, H., and Harjumaa, M. Towards deeper understanding of persuasion in software and information systems. In *First International Conference on Advances in Computer-Human Interaction*, IEEE (2008), 200–205.

26. Pirolli, P., and Card, S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, vol. 2005 (2005), 2–4.

27. Rodgers, R., and Hunter, J. The discard of study evidence by literature reviewers. *The Journal of Applied Behavioral Science 30*, 3 (1994), 329.

28. Savikhin, A., Maciejewski, R., and Ebert, D. Applied visual analytics for economic decision-making. In *IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST'08* (2008), 107–114.

29. Simon, H. Rationality as process and as product of thought. *The American Economic Review 68*, 2 (1978), 1–16.

30. Stangor, C., Sechrist, G., and Jost, J. Social influence and intergroup beliefs: The role of perceived social consensus. *Social Influence: Direct and Indirect Processes* (2001), 235–252.

31. Stasko, J., Gorg, C., and Liu, Z. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization 7*, 2 (2008), 118–132.

32. Sunstein, C., and Thaler, R. Libertarian paternalism is not an oxymoron. *U. Chi. L. Rev. 70* (2003), 1159.

33. Thaler, R., and Benartzi, S. Save More Tomorrow: using behavioral economics to increase employee saving. *Journal of political Economy* (2004), 164–187.

34. Thaler, R., and Sunstein, C. *Nudge: Improving decisions about health, wealth, and happiness*. Yale Univ Pr, 2008.

35. Thomas, J., and Cook, K. Illuminating the path: The research and development agenda for visual analytics. *IEEE Computer Society* (2005).

36. Wason, P. Reasoning about a rule. *The Quarterly Journal of Experimental Psychology 20*, 3 (1968), 273–281.

37. Wright, W., Schroh, D., Proulx, P., Skaburskis, A., and Cort, B. The Sandbox for analysis: concepts and methods. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM New York, NY, USA (2006), 801–810.

38. Yang, D., Rundensteiner, E., and Ward, M. Nugget discovery in visual exploration environments by query consolidation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ACM New York, NY, USA (2007), 603–612.

39. Experimental data. `http://removed_for_anonymity`.