On the Problem of Predicting Real World Characteristics from Virtual Worlds

Muhammad Aurangzeb Ahmad, Cuihua Shen, Jaideep Srivastava and Noshir Contractor

Abstract Availability of massive amounts of data about the social and behavioral characteristics of a large subset of the population opens up new possibilities that allow researchers to not only observe people's behaviors in a natural, rather than artificial, environment but also conduct predictive modeling of those behaviors and characteristics. Thus an emerging area of study is the prediction of real world characteristics and behaviors of people in the offline or "real" world based on their behaviors in the online virtual worlds. We explore the challenges and opportunities in the emerging field of prediction of real world characteristics based on people's virtual world characteristics, i.e., what are the major paradigms in this field, what are the limitations in current predictive models, limitations in terms of generalizability, etc. Lastly, we also address the future challenges and avenues of research in this area.

1 Introduction

> When the number of factors coming into play in a phenomenological complex is too large scientific method in most cases fails.
>
> —Albert Einstein in *Out of my later years* [12]

Although somewhat simple but it would not be a great exaggeration to state that the Sciences consist of two important parts—the descriptive and the predictive. Yet, for the most part, predicting human behaviors has been a more challenging task than describing them. The reasons behind this are twofold: first, until very recently, it has

M. A. Ahmad (✉) · J. Srivastava
University of Minnesota, Minneapolis, MN, USA
e-mail: mahmad@cs.umn.edu

C. Shen
University of Texas at Dallas, Richardson, TX, USA
e-mail: shencuihua@gmail.com

J. Srivastava
e-mail: srivasta@cs.umn.edu

N. Contractor
Northwestern University, Evanston, IL, USA
e-mail: nosh@northwestern.edu

M. A. Ahmad et al. (eds.), *Predicting Real World Behaviors from Virtual World Data,*
Springer Proceedings in Complexity, DOI 10.1007/978-3-319-07142-8_1,

been extremely difficult and costly to collect accurate behavioral data about human beings in large amounts. Second, human behaviors, the very phenomena that the Social Sciences are studying, are much more complex as compared to the subjects of study of natural sciences. The last 15 years have seen an explosion of digital trace data that can be collected about human behaviors and thus it has also enabled us to not only answer traditional questions in social sciences but also ask new questions, which have not been possible to be addressed because of lack of data in the past. This has given rise to the field of Computational Social Science [19] and researchers in the area have likened the current state of affairs in this field to the circumstances that gave rise to Cognitive Science in the 1950s with the emergence of new types of inter-disciplinary collaborations and availability of new types of data. The emergence of this field offers us new opportunities and challenges with respect to new methods since traditional methods of data gathering and analysis may not scale well or it could be the case that the older methods were not designed with the new type of data in mind.

Massive online games (MOGs) or massively multiplayer online (MMO) games are online games that are characterized by shared persistent online environments where hundreds of thousands, and in some cases, millions of users can simultaneously be part of the virtual environment and interact with one another and also with the environment. Well-known examples of MMOs include World of Warcraft (WoW), EVE Online, EverQuest, EveryQuest II, Star Wars: Knights of the Republic (SWTOR), etc. Given that a large subset of the population in these environments actually spends a significant amount of time in these spaces, the questions arises: do people still have the same persona when they are playing online and how much of their offline persona do they bring online when they are playing in these environments? Hence we would like to know if it is possible to use data from virtual worlds to predict about the "real" world characteristics of players e.g., gender, age, location, deviance, personality, ideology (political), etc. We address these and related questions in this chapter. We consider the application domain and how these virtual worlds map to their real world counterparts as well as where the mapping fails. Additionally, we discuss issues related to generalization of results obtained from studying these environments, how data quality impacts analysis, data augmentation, different approaches that can be used for data modeling, etc. Many of the insights in this chapter are based on our work in the Virtual World Observatory (VWO) project as well as our work in the gaming industry.

2 The Mapping Principle

The Mapping Principle [26] starts with the simple observation that some human behaviors in virtual spaces are often quite similar to their counterparts in the offline flesh and blood world. Dmitri Williams argues that the Mapping Principle cannot be taken for granted in the virtual places, at least not at this point in development of virtual worlds, and has to be established on a case-by-case basis. From this observation

it becomes obvious that not all virtual worlds and virtual behaviors map to the offline world. Williams notes that successful mapping in the virtual worlds offers certain advantages, which were not available to research in the social sciences before access to such types of data became available recently. For example, many of the social structures that one observes in the virtual world have offline counterparts, but data collection of such structures in the offline world is often extremely limited by resource constraints. These also offer possibilities with respect to studying human behaviors at different levels of granularity, i.e., at the individual, group, and society levels.

Williams also notes the problem of representation in studying virtual worlds, i.e., different virtual worlds may represent a user in different ways and thus this may make the task of mapping from one virtual environment to another virtual environment quite complicated. This means that we can take it for granted that one virtual world is similar to another virtual world and thus some sort of mapping has to be also established from one virtual world to another world. Even within same type of virtual worlds, different types of social environments can elicit different types of behaviors and should be part and parcel of standard analysis. For example, Player versus Player (PvP) and Player versus Environment (PvE) types of gameplay are quite different with respect to eliciting cooperative or aggressive behaviors from their respective players. Lastly, another complication lies in the fact that even an isolated virtual world is not really isolated since human beings by their very nature bring certain cognitive biases with them. This is clearly evident in the "Proteus Effect" [27] where people's behaviors in the virtual worlds are greatly dictated by the characteristics of their avatars regardless of whether the avatars are similar to their offline characteristics or not.

3 Theory-Driven versus Data-Driven Paradigms

The history of the various sciences is characterized by two different yet complimentary approaches to hypothesis formation and testing: the theory-driven approach and the data-driven approach. We consider these paradigms within the context of social sciences. Given a certain phenomenon, the theory-driven approach takes a social science theory or a set of theories as the guiding principle for data collection and exploration purposes. The type of data that are collected is that guided by or even constrained by theory. Thus consider the phenomenon of friendship and the motivations behind why people form relationships. Social science theories would suggest factors related to homophily, proximity, personality, etc. that would account for the formation of friendship relationships [18, 21]. The data-driven approach, on the other hand, takes a theory-agnostic view with respect to data collection and analysis. While all types of data are collected by considering technological, ethical, and resource constraints within the data-driven paradigm, the analysis is guided by traditional computational concerns like scalability, dimensionality reduction for tractable analysis, information theoretic measures for determining data relevance, etc [15].

It is important to note here that the two approaches are not disjoint; theory is always informed by data. What we mean here by the data-driven approach is that, more often than not, this approach does not presuppose a theory vis-a-vis data collection and its interpretation. Big data add a new dimension to these two approaches. One drawback of the data driven approach is that one can end up with a black box where the variables that one may infer with respect to explaining a social phenomenon may not make sense from a theory perspective. In the ideal case, the mismatch between the theory and the data can reveal new insights with respect to the underlying phenomenon and thus may even help update the theory. In the worst case, these new factors may be treated as epiphenomenon from the perspective of social scientists who want models that are always tied to some social or psychological explanation. In most cases, however, the data-driven approach can indeed be used as a source of feedback to the theory-driven approach where the latter is no longer limited by traditional methods of data collection.

A number of studies have been done in this area to demonstrate the efficacy of the mutually reinforcing nature of these two approaches. Thus, Ahmad et al. [19] addressed the problem of detecting Gold Farmers in MMOs where they used traditional domain rich social science approaches for data analysis and then augmenting that approach with machine learning techniques for feature construction and model building. Similarly Borbora et al. [7] describe the use of both of these approaches for determining churners in MMOs. One set of features are chosen based on literature that describes psychological factors and motivations for engagement and play and another set is chosen based on information theoretic measure like information gain. The main conclusion from their work is that a combination of both these approaches is most accurate for prediction tasks. The final choice of the feature sets can also shed some light on the psychological factors and motivations, which may have been missed by theory.

4 Limitations and Methodological Issues

While Big Data offers us opportunities to gain insights into human psyche and social behaviors, it should not be taken to be the be all and end all of studying human behaviors. There are a number of methodological issues with respect to using log data from virtual worlds and MGOs that preclude one from making certain conclusions about human behaviors. Additionally there are certain limitations, such as data availability, collecting missing data after the fact, issues related to generalization across environments and generalization in the real world, issues related to self-reported data, and how incorrect mapping can actually lead to incorrect conclusions. We discuss each of these issues in some detail in the following subsections and also offer solutions to partially mitigate these problems.

4.1 Data Quality

Game logs and other databases in case of virtual worlds can provide a quite comprehensive view of a user's characteristics and activities within a virtual environment, but they do not always capture everything about the environment being studied, e.g., psychological motivations, reaction to outside events such as divorce, drug abuse, religious conversion, etc. Thus, even in the ideal case where the data logs are recording everything, it is still the case that *everything is not everything*. This particular issue can be partially mitigated by complementing data from other sources like surveys, which we discuss in detail in Sect. 4.3.

In case of MMOs and other virtual environments, data can be divided into two main types: user characteristic data and user activity data. The user characteristic data consist of demographic characteristics, which are immutable in almost all cases like gender, age (which progresses at a constant rate) and semi-immutable characteristics constitute the characteristics that a user adopts for her virtual character, e.g., the character's gender, race, class, etc. The user activity data, on the other hand, consist of activities that a user performs in the gaming environment. While the user activity data can be said to capture all of the relevant that a person performs in a game without any filter, the same cannot be said about the user characteristic data since it is manually entered by the user. Thus it is possible for people to lie about their gender, age, location, etc. when they register to play a game. As an example, we consider the user data from EverQuest II described above where we observed that a subset of deviant players known as gold farmers always specified their location as either Alabama or Antarctica. The reason for choosing Alabama is likely because it is the first option to choose for a place of residence among the states in the USA. Such cases of obvious misdirect by the users have to be excluded from the analysis. Similar misdirects with respect to a person's age and gender can also happen and can be partially captured if additional sources of information for the same type of data are available.

When the game company is storing most activity data from its users, a number of data-related issues still have to be addressed. First, given the cost and storage and retrieval of large amounts of data the level of granularity at which the data are being saved can vary greatly. For example, should the telemetry information (location of a player character) in an MMO be saved at every microsecond or should it be saved whenever a player interacts with another player or a nonplayer character (NPC)? Notice that the former scheme is likely to generate massive amounts of telemetry data for games like WoW, which have millions of active players. Third, game logs and the corresponding database schemas are not designed with specific questions in mind. This can lead to poor performance with respect to data querying and retrieval.

4.2 Generalization

Issues related to generalization have been part and parcel of social sciences since their inception [13]. Traditionally, it has been extremely difficult to do experiments

in the social science outside of the laboratory setting, which limits the size of the participant pool because of practical resource constraints. Generalization was thus dependent upon having enough observations over a number of populations which are varied enough [13]. There are also practical and ethical constraints with respect to manipulating human beings for the purpose of gathering data. In virtual worlds, however, since the cost of participation and making changes to the infrastructure is quite low as compared to their real world counterparts, it is possible to do things like AB testing on a massive scale. While virtual worlds offer a way to add at least partial intervention for data collection, they still leave open the question of replication across different virtual worlds. Consequently, the question of generalization *across* virtual worlds is still an unaddressed question as we shall argue below.

One of the principle issues with respect to studying MMOs is the lack of publically available datasets. Almost all the work that has been published in this area is either by scraping data from MMO websites or by researchers who get access to the datasets by working with the organizations that created the MMOs. As a consequence, access to the dataset is limited because of Non-Disclosure Agreements and confidentiality agreements. There are also legitimate liability and privacy concerns from the perspective of the gaming companies, which preclude sharing of the datasets as the Netflix data de-anonymization debacle demonstrates [20]. Thus as a result of these limitations, replication of results is limited and generalization of results is even more severely limited because researchers in general do not have access to multiple datasets. This can be especially problematic when researchers compare real world social interactions and those in the virtual world. The main epistemological risk is that generalizing results from one MMO or by comparing just one MMO with real world data.

A representative case of this issue is the work by Johnson et al. [16] on team formation in guilds by using WoW Data and the follow-up work by Ahmad et al. [3] on the replication of their results using EverQuest II data. Johnson et al. compared team data from guilds in WoW and street gangs in Los Angeles and proposed a model that can replicate the team size distribution in both these datasets. From their observations they concluded that a single mechanism is responsible for team formation in both offline and online settings. Ahmad et al. used the same model that was proposed by Johnson et al. and applied it to the guilds in EverQuest II. The results that were obtained by Ahmad et al. [3] were almost opposite to the results obtained by Johnson et al. The original model was predicated on the fact that the main driving force in team formation is the maximization of skillset but Ahmad et al. [3] were able to replicate these results by using a model that favors homophily, a mechanism explicitly ruled out by the original paper [16].

A number of lessons can be learned from this cautionary tale. One possibility is that it could be the case that the results from the WoW study are generalizable to the offline world but not from EverQuest II. The second possibility is that the convergence of results between WoW and the Los Angeles gangs' dataset was a fluke and the results are not generalizable from the study. The third possibility is that it is the EverQuest II dataset that corresponds well with the gang dataset. Yet another possibility is that there are yet undiscovered common generative mechanisms that

describe team formation for all the three cases. Lastly, we note that the fifth possibility is that the gang in the offline world and the guilds in MMOs do not constitute a good mapping to draw any meaningful conclusions about the other. Regardless of which of these possibilities are correct, one observation that is common across all of these scenarios is that there are serious issues with respect to generalization unless data are employed to replicate the same test across multiple MMOs.

Even if we can do the mapping exercise correctly, it does not guarantee generalization. Thus claims regarding generalization based on one or two datasets should either come with the appropriate disclaimers or should be considered with extreme caution.

4.3 Surveys, Perception, and Truth

As described in the previous sections, even a perfect log database does not capture everything about a player's behavior. Information in the logs can be augmented with additional information by doing surveys of the players for whom the log data are already available. Additional sources of information can be added to game logs via social media websites like Facebook and Twitter, as well as online forums and communities, with the permission of the users. Information from these social networks can be used to augment in-world social network information between the users. Thus, veracity of certain types of information can be established to a greater degree if the same information is coming from multiple sources. For example, it could be the case a person is gender bending not just their virtual character but also reporting their real gender incorrectly in other environments. Thus in the EverQuest II data we also discovered instances where the real world gender reported by people on surveys is different from the real world gender that they reported when they signed up to play the game. Without an additional third source of information there is no way to disambiguate the gender of these players and thus these have to be left out of the analysis.

Out of these sources surveys can be a rich source of data to augment log data but survey data comes with its own set of unique problems. A number of studies spanning decades have shown that people do not always tell the truth on surveys [5, 23], either consciously or unconsciously, which calls the veracity of survey data partially into question. To illustrate this point, let us consider the example of amount of time a player spends playing games versus the amount of time that a player reports that he or she spends playing a game. In a study conducted by Williams et al. [24] it was discovered that both men and women underreport the time that they spend playing video games. Women actually underreport more as compared to men. The self-report information on the amount of time spent playing the game was determined from survey data, which was in turn linked to the game log data to infer the discrepancy between the reported hours and the actual number of hours spent. These results can be interpreted in a number of ways: it may be the case that people in general do not have a good handle on how much time they spend playing games. Or, people like to

underreport because of some social stigma associated with spending too much time playing video games [25]. In both the cases, however, the result is the same—the time discrepancy remains. Thus any report which uses complementary survey data and limited options to add corrective measures should come with the appropriate disclaimers.

4.4 Mismapping from Virtual Worlds to the Real World

The Mapping Principle can only work if one knows what one is mapping to in the virtual world. In many cases, the virtual and the real will have many similarities and even share the same terminologies but it would be unwise to assume the two are equal. More often than not, one runs into examples where finding maps from one domain to the other but the conclusions from such studies are lacking because the mapping is not done properly. We refer to such cases as Mismapping. A common scenario where this happens is what the researchers do not invest enough time and effort in gaining expertise in the online domain (the virtual world or the MMO) that they are studying. Examples of such cases include misidentifying attributes of player characters [24] or misunderstanding of game mechanics [16]. The net effect of Mismapping is the tendency to make conclusions about the phenomenon.

Mismapping can, however, be easily avoided by investing some time and effort in actually playing the game in case of MMOs or spending some time getting familiar with the environment in the case of other virtual worlds. Usually spending 20 h in an environment seems to be a good guiding number for acquiring the minimal levels of expertise in such content-rich environments [26]. We would go as far as to recommend that conferences and journals should have a mandatory policy for requiring at least one author for any paper to have spent at least about a minimum of prespecified time in the virtual world that they are studying in order to avoid Mismapping. To illustrate this issue we again turn to the generalized team formation model of Johnson et al. [16]. In the paper, they assume that the guilds in WoW are analogous to ethnic groups in the street gangs in Los Angeles. This assumption is unwarranted because ethnicity based groups do not bear much resemblance to guilds in MMOs beyond a superficial level. Even if one were to make the argument admissible the organization of guilds in MMOs can range from something akin to playgroups to military style organizations [24] and would thus preclude any meaningful mapping between the two.

5 Case Studies: Virtual to Real World Mappings

There are a number of phenomena that are observed in the virtual world, which have their counterparts in offline worlds so that one can clearly observe similar types of behaviors in both the settings. The Mapping Principle [26] can be used to determine in which cases this mapping can be applied to these phenomena. In this section, we

examine a few case studies where the Mapping Principle has been applied. We also investigate where mapping from the virtual to the real succeeds, and where it fails and the reasons for the failure.

5.1 Case Study: Virtual Economy

Economies of MGOs were one of the first phenomena to be observed in detail when such data started to become available. Both online and offline economies are characterized by finite resources and availability of resources which are driven by supply and demand. Certain games like EVE Online have actually hired professional economist to design the economy systems within [14]. A number of pioneering studies of the virtual worlds was done by Edward Castronova on the economies of virtual worlds [8–10]. It was observed that the virtual world economies exhibit many patterns that are observed in their real world counterparts. An added advantage of using virtual worlds as test beds for studying economic behavior is that they offer a way to simultaneously study both the microeconomic as well as macroeconomic behaviors of people. In the offline world it is not possible to do so because of privacy concerns and also because of the fact that data related to economic activities are spread over a vast number of sources. Virtual worlds also offer us a way to do AB testing in a manner that is not possible in the offline world. Even different economic systems with variants can be tried out to determine how people would act under different economic affordances and constraints.

The economies of the virtual worlds have many characteristics that one associates with a functioning economy: buying and selling of goods, creation and destruction of goods, banking, a parallel shadow economy in the form of gold farmers [1, 17], etc. Castronova et al. [10] studied the aggregate economic behaviors of players in EQ2 and compared the behaviors with real world economies and what theories of macroeconomic behaviors say about how humans should behave economically. The study reveal that aggregate economic measures like the GDP, price level, money supply, and inflation behaved in exactly the same way that their counter parts behave in the offline world, e.g., when the prices go up then demand decreases, when the population decreases which causes a decrease in demand then the prices go down. A natural experiment also occurred in their dataset where a new server was introduced in the game and the aggregate economic behavior of the new server quickly mimicked the aggregate behavior of the already existing servers.

Given the positive results the researchers suggested additional possibilities for future which can even have real world implications: researchers have traditionally used simulations to study the effect of changes in economic systems but virtual worlds can offer an even more realistic way in which people's reaction to an economic policy can be gauged. So before that change is introduced in the real world, policy-makers can try out such changes in the virtual world. The risks associated with implementing a policy in the virtual world with negative consequences would be far less as compared to implementing it in the offline world. Another intriguing part of

Fig. 1 The corrupted blood incident in WoW

the virtual economy is the clandestine black market economy that exists just like its offline counterpart [1, 17]. The main difference, however, is that in the offline world it is next to impossible to collect data about black market activities but such constraints are less severe in the virtual worlds [1]. Previous research has even shown that the behaviors of clandestine actors in the virtual worlds are very similar to their counterparts in the offline worlds [17].

5.2 Case Study: Epidemiology

The Corrupted Blood incident is a famous "global" event in WoW where a large number of players were affected by an unplanned in-game event [6] and seemed to offer potentially interesting insights to researchers of epidemiology. On September 13, 2005, Blizzard, the creators of WoW (Fig. 1), introduced a new dungeon instance into the game with a new boss associated with the dungeon. The new boss had a spell (Corrupted Blood) that acted like an infectious disease. Once a player was infected they could transmit the disease to other players and even NPCs if they were close enough. The spell damaged the "health" of the other characters over time. Blizzard

had created the spell to be confined to the dungeon instance where it was introduced. But the unintended consequence of the spell is that it started to spread like a plague and eventually infected a large percentage of players in a number of servers [6]. A number of researchers in the field of epidemiology noted the similarities between how people reacted to Corrupted Blood in WoW and how people react to epidemics in the real world. Thus, Balicer [6] noted in one of the first papers on the subject, "a platform for studying the dissemination of infectious diseases, and as a testing ground for novel interventions to control emerging communicable diseases."

Beyond the obvious similarities between real world epidemics and the contagious nature of Corrupted Blood, researchers noted a number of other similarities: the Corrupted Blood originated in a remote, uninhabited region in WoW. It was carried by travelers to larger regions (Fig. 1) and also by players who were actively fleeing the main centers of the plague, the hosts for the plague were both human and animal [6]. In these respects some regarded that it had similarities to the Avian flu virus [6]. The similarities, however, end here and we should examine the efficacy of using virtual worlds to study these phenomena in the historical context. Earlier work in the field of Social Simulation focused on simulating influence and the spread of outbreaks but in a purely simulation-based framework where human agents are substituted for artificial agents [26]. Thus the main limitation of this approach was that certain assumptions that were made about human behavior like rationality or even bounded rationality were not borne out by psychological studies of human behavior [28]. The excitement of the research community sprung from the fact that these virtual environments represented spaces where one did not have to rely on virtual agents anymore and collect real data about human reactions to contagious diseases without any harm being done to humans.

On the other hand, the lack of liability on the part of the human beings is *also* a problem with respect to correct modeling of the phenomenon. The maximum penalty in WoW for contracting plagues is the death of the player's character, which can be regenerated. The player would at most lose some of the virtual resources acquired, but in the real world, a person who loses her life has no hope of ever being resurrected. People are likely to behave very differently when their lives are at stake. In conclusion, while there are a number of similarities between the two environments and the Mapping Principle seems to work well at first glance but the most important aspects of the two environments do not map and thus preclude us from making many useful conclusions about the real world in this case.

5.3 Case Study: Deviant Clandestine Behaviors

The phenomenon of deviancy is socially constructed. Thus, what may be considered deviancy in one context and culture may not be considered deviancy in another [1]. While deviant behaviors may or may not be clandestine, an important subset of such behavior is in fact clandestine in nature and these have mainly evaded analysis because of lack of data. It should be noted that clandestine activities may or may not be

illicit since what constitutes illicit is also socially constructed. The common factor in the study of clandestine activities is that there is some effort by the parties involved to hide their activities. The Webster's dictionary defines clandestine activities as those activities which are "kept secret or done secretively, esp. because illicit." In the larger scheme of things it is extremely difficult to study deviant clandestine behaviors because of not just privacy implications but also the extreme difficulty to obtain data about criminal outfits. Here the Mapping Principle may be of benefit for, i.e., while the severity of deviant behaviors in the virtual worlds may be less, the constraints and affordances under which people operate in the two environments. With this observation in mind we consider various operationalizations of deviancy and clandestine behaviors in MMOs and what can be learned from them.

We revisit the WoW guilds example and the street gangs of Los Angeles [16]. The former represents a case of virtual organizations and the latter represents a possibly criminal organization, which has at least some elements of being clandestine in nature. In this case as well one cannot map the two environments because guilds in WoW are not clandestine in nature and street gangs are also semi-public in nature. To be fair, the authors of the paper do not make any claims regarding the deviant or the clandestine nature of the datasets. Thus, the affordances and the constraints of a clandestine environment do not map to the virtual world setting in the case of guilds. We now consider the example of gold farmers in MMOs; these are players who are involved in activities that are considered "illegal" by game administrators as they disrupt the economic balance of the game as well as challenge the assumption of the game as a meritocracy [1, 17]. Game administrators seek to actively ban gold farmers, and as a result the gold farmers change their behaviors to avoid detection [1]. This in turn forces the game administrators to update their strategies to catch gold farmers and then the cycle continues.

The gold farmer detection avoidance detection behavior is not restricted to simple behaviors but manifests in more complex ways, e.g., the gold farmers would extend their supply chains in order to put multiple layers of obfuscation between them and the game admins [1]. Gold farmers also avoid interacting with one another in modalities—for example, trust exists between pairs of gold farmers, which may indicate the presence of a strong relationship, but they interact with one another via proxies [17]. It has been noted that the avoidance detection techniques employed by the gold farmers [1], the thinning out of social networks as a tactic [17], etc., are very similar to how drug dealers operate. Thus a study by Keegan et al. [17] observed that there are a number of similarities between the social networks of gold farmers and the social networks of drug dealers, a dataset collected by the Calgary police department [17]. Keegan et al. [17] argue that the conditions under which the gold farmers operate are similar to the conditions under which drug dealers have to operate. In other words, in both cases the commodity being sold is limited in quantity, there is an active campaign by the authorities to clamp down on its distribution, there is a long supply line between the source and the distribution of the goods, important actors in the network placing themselves in noncentral positions where they are difficult to detect, etc.

The comparison by Keegan et al. [17] between the social networks of gold farmers and drug dealers revealed that these networks are much more similar to each other as compared to other social networks. At the surface, the two networks seem quite divergent but the presence of similar constraints and affordances reveal the presence of common generative mechanisms. This opens the possibility to study clandestine networks and behaviors in sufficient detail which is not possible in the offline world. We may be able to learn something about the offline counterparts of the virtual world deviant actors given a sufficiently careful study. A word of caution is however in order, it may be the case that the mapping in between the two environments works in the case of Keegan et al. [17] but the results may or may not be generalizable. Additionally, it is premature to base any policies based on the gold farmer and drug dealer comparison but one can still learn insights regarding their respective social networks given that the virtual world data are much rich in content [26].

5.4 Case Study: Mentoring

Mentoring is a phenomenon where a more experienced person in a given field serves as an adviser or a trainer to a less trained person [2]. There is an extensive amount of literature on mentoring in organizational theory [2] and small group research [2]. However, studies focusing on mentoring networks have been very limited and to date less than half a dozen such studies have been published [2, 4]. This is mainly owing to the fact that historically it has been extremely difficult in collecting mentoring data in a network setting. Prior work in this area can be divided into mentoring networks in the field of psychology and mentoring networks in MMOs. Temporal data for the psychology network is not available so that a meaningful comparison cannot be made between the two networks. Ahmad et al. [4] and Huffaker et al. [2] have done a series of studies on mentoring in MMOs. One of the things that they discovered was that the mentoring networks were different from most other social networks [4]. This of course raises questions regarding generalizability of not only social networks in MMOs but also domain-specific social networks. There are a number of motivations that people have with respect to mentoring others. The literature notes that people mentor because of instrumental reasons, friendship, organizational obligations, or a sense of paying it forward [2].

Starting with a set of features that characterized mentoring activities in the MMO EverQuest II, Huffaker et al. [4] observed that the most optimal clusters that they obtained corresponded to the four categories that are described in the literature. This observation also established at least some limited form of mapping between the virtual world and the offline world with respect to mentoring. The virtual world data, however, is also complemented with network information between the mentors and their apprentices. Huffaker et al. [4] observed that the various clusters of mentors also have different structural characteristics: instrumental mentors have a much higher value for closeness centrality as compared to other types of mentors, even though the instrumental mentors have a low value for mentoring. The explanation here is

that the instrumental players are more focused on their own achievement but since they do not confine themselves to a small circle of friends their network spread out more. Also players who are focused on their virtual organizations or guilds have much higher clustering coefficients as compared to other types of mentors. Again this is expected, because mentors with this form of motivation naturally have greater engagement with other people. While most of these observations agree with what the researchers in the organizational studies' literature have suspected, the main lesson here is that since virtual worlds can offer us ways to learn about the real world and even test hypothesis which may not be possible in the offline world.

6 Predictive Modeling from Virtual World(s) to Real World(s)

In Sect. 1.3 we discussed the issues inherent in comparing the data-driven versus theory-driven paradigm. We note that the same debate has important implications not just for hypothesis formation and testing but also for predictive modeling. Consider a prediction task where the goal is to build models which also have explanatory power in terms of the features or the variables used. As an expositionary example, consider a behavioral prediction problem. There are two main ways in which such models can be constructed. We first consider the **theory-driven** paradigm: one can start with existing social science theories of human behavior and select or construct variables based on what the theories say about human behavior in that context. This can of course be augmented with new features which may appear in a new domain but which nonetheless fall with the purview of existing theory. Existing theories are also used to form hypothesis regarding the domain, these in turn are tested given the data. This is illustrated in Fig. 2 which also shows the alternative paradigm, the **data-driven** paradigm. In the data-driven approach one starts without any preconceived notions of how theory should drive model building. Models are built by features or variables which can be selected based on criteria like Information Gain, PCA, or other similar criteria. The process of model building itself is iterative in nature which may include things like parameter tuning, oversampling depending upon the distribution of classes, label propagation, etc.

It is important to note here that the two approaches are not exclusive in practice. There is always some element of data driving the direction of model building even in the case of theory building. And data-driven approaches may select different variables depending upon the domain where it is being applied even if feature sets and models themselves are not dependent upon the theory in the initial stages. However, treating their approaches as separate makes sense from a practical perspective. Consider the scenario given in Fig. 2 where the two approaches are being used to solve the same problem: four outcomes are possible. In the first scenario, the results from both the theory-driven and the data-driven approach agree, which implies that the data-driven results augment what the theory states. Thus, additional iterations of the model building process are not required. In the second scenario, one gets good results from the theory-driven approach but not from the data-driven approach. This would be the

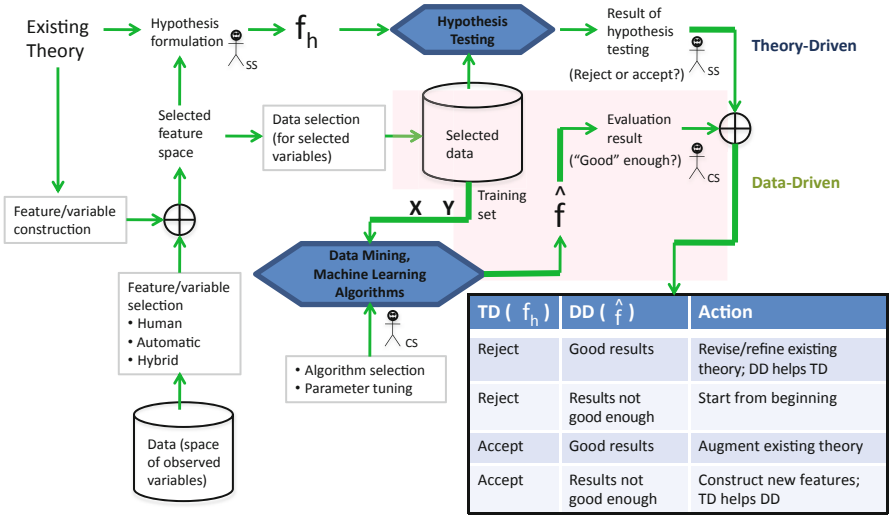

Fig. 2 Predictive models with social science theories

case where the data-driven models need to be updated based on insights from the theory. This may involve construction of new features from theory to augment model building. In the third scenario, one gets good results from the data-driven approach but not from the theory-driven approach. This is interesting because it implies that existing theory actually may have missed something and needs to be updated. This may involve updating existing theory with new explanatory variables, although the process may involve complications as we shall describe below. The last scenario is the trivial case where both the theory-driven and the data-driven approaches give weak results and the whole process of model building has to be revised from the beginning.

The particular approaches of actual model building for predicting real world from virtual world can vary depending upon the actual characteristics. The space of real world characteristics can be divided into ascribed characteristics and acquired characteristics [4]. Ascribed characteristics are the characteristics that are immutable in almost all the cases, such as gender, age (changes at a constant rate), personality-type (which changes rarely) [4], aptitude, intelligence quotient, emotional quotient, etc. Notice that these in turn can be divided into two subtypes: demographic characteristics and psychometric characteristics. The second class of characteristics is called acquired characteristics; these include education, location, vocation, etc. These are the characteristics of a person which can change over time thus a person who has a high-school diploma can get a bachelor's degree and have a different education status. For each of these variable types the choice of feature depends upon the domain as well as the variable that needs to be predicted. Consider the case of real world gender. Prior work has shown that while gender-bending is more common amongst

men in virtual worlds [11], the gender of a person's primary character, i.e., the character with which they spent the most time playing is a good indicator of their real world gender. Similarly the choice of a character's virtual race and class seems to be a good indicator of their political and ideological leanings [26]. These observations should not be surprising as people do tend to take a part of them in the virtual worlds.

While the data-driven approaches are usually complementary to the theory-driven approaches, there can be cases where data-driven models may lack in explanatory power. Consider a dataset for which classifiers, like certain SVM [15, 22], may have high values for precision and recall but which are essentially black boxes that cannot shed much light onto why the model works. Now let us suppose that a more explanatory model like JRIP or Decision Tree [15] do not work well with a dataset and gives a low value for precision and recall. The dilemma for the computational social scientist here is to find a balance between explanatory models versus models that predict well. Explanatory models sometimes come at the cost of performance and vice versa. The solution to this dilemma then depends upon the application and the domain where the models are being applied. Thus if the goal is prediction then the black box nature of the model would be irrelevant but if the goal is to explain why the model works and how the various parts of the model fit together then one is left with no choice but to use the more explanation rich model at the expense of performance. While this tradeoff is not always needed, but it happens often enough that it is of note.

7 Conclusions

Virtual worlds have their origins in multi-user-dungeons in the late 1980s, which allowed multiple players to share the same gaming environment [1]. From these early environments they have evolved to become environments where millions of people can share these online persistent worlds. There are more than 400 million people who play MMO games in one form or another. This number itself represents 5.7 % of the population of the world and this number is likely to grow in the future. This implies the opportunities for further research in this area also abound. One can also think of the virtual world environments as natural experiments given the practical and logistical difficulties in collecting data about people in large assemblages. Given that many people who partake in virtual worlds do in fact spend significant time and effort in these virtual worlds, it should not be surprising that people's real world characteristics affect their gameplay and behaviors online. The Mapping Principle establishes mapping between virtual environments and their real world counterparts given that certain conditions are met. There are several examples where mapping may have been done based on insufficient similarities and thus prematurely mapped, which we caution as an example of how not to do computational social science. Another important point to keep in mind with respect to the Mapping Principle is that even if one can do the mapping exercise correctly it does not guarantee generalization. Results may not generalize to other MMOs let alone to the offline world.

The process of model building for predicting real world characteristics is iterative in nature and one can adopt data-driven as well as theory-driven approaches. These paradigms are complementary in nature and describe two different ways to conduct computational social science. One possible issue that can be problematic is the issue of tradeoff between explanatory powers of models versus their performance. In such scenarios, the decision regarding which models to use depends upon the demands of the application. The choice of variables that are used usually depends upon the domain and the context.

Acknowledgments Special thanks to Mushtaq Ahmad Mirza and Khalida Parveen for being there.

References

1. Ahmad, M.A., Keegan, B., Srivastava, J., Williams, D., Contractor, N.: Mining for gold farmers: Automatic detection of deviant players in MMOGs. In: Computational Science and Engineering, 2009. CSE'09. International Conference on (vol. 4, pp. 340–345). IEEE (2009, August)
2. Ahmad, M.A., Huffaker, D., Wang, J., Treem, J., Kumar, D., Poole, M.S., Srivastava, J.: The many faces of mentoring in an MMORPG. In: Social Computing (SocialCom), 2010 IEEE Second International Conference on (pp. 270–275) IEEE (2010, August)
3. Ahmad, M.A., Borbora, Z., Shen, C., Srivastava, J., Williams, D.: Guild play in MMOGs: rethinking common group dynamics models. In: Datta, A., Shulman, S. W., Zheng, B., Lin, S.-De, Sun, A., Lim, Ee-P. (eds.) Social Informatics, pp. 145–152. Springer Berlin Heidelberg (2011)
4. Ahmad, M.A., Ahmed, I., Srivastava, J., Poole, M.S.: Trust me, i'm an expert: Trust, homophily and expertise in MMOS. In: Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom) (pp. 882–887). IEEE ($62011, October)
5. Babbie, E.R.: Survey research methods. Wadsworth, Belmont (1990)
6. Balicer, R. D.: Modeling infectious diseases dissemination through online role-playing games. Epidemiology **18**(2), 260–261 (2007)
7. Borbora, Z., Srivastava, J., Hsu, K.W., Williams, D.: Churn prediction in MMORPGS using player motivation theories and an ensemble approach. In: Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom) (pp. 157–164). IEEE (2011, October)
8. Castronova, E.: Synthetic worlds: The business and culture of online games. University of Chicago Press, Chicago (2005)
9. Castronova, E.: On the research value of large games natural experiments in Norrath and Camelot. Games and Cult. **1**(2), 163–186 (2006)
10. Castronova, E., Williams, D., Shen, C., Ratan, R., Xiong, L., Huang, Y., Keegan, B.: As real as real? Macroeconomic behavior in a large-scale virtual world. New Media Soc. **11**(5), 685–707 (2009)
11. Davis, D.Z.: Gendered performance in virtual environments. Media Disparity: A Gender Battleground (2013): 133.
12. Einstein, A.: Out of my later years. Citadel Press, Secaucus (1956)
13. Fiske, D.W., Shweder R.A. (eds). Metatheory in social science: Pluralisms and subjectivities. University of Chicago Press, Chicago (1986)
14. Games, C.C.P.: Eve online. World Wide Web. http://www.eve-online.com (2003). Accessed 5 May 2014

15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
16. Johnson, N.F., Xu, C., Zhao, Z., Ducheneaut, N., Yee, N., Tita, G., Hui, P.M.: Human group formation in online guilds and offline gangs driven by a common team dynamic. Phys. Rev. E. **79**(6), 066117 (2009)
17. Keegan, B., Ahmad, M.A., Williams, D., Srivastava, J., Contractor, N.: Dark gold: Statistical properties of clandestine networks in massively multiplayer online games. In: Social Computing (SocialCom), 2010 IEEE Second International Conference on (pp. 201–208). IEEE (2010, August)
18. Kandel, D.B.: Homophily, selection, and socialization in adolescent friendships. Am. J. Sociol. 427–436 (1978).
19. Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. Van Alstyne, M.: Life in the network: The coming age of computational social science. Science (New York) **323**(5915), 721 (2009)
20. Narayanan, A., Shmatikov, V.: How to break anonymity of the Netflix prize data set. The University of Texas at Austin, Austin (2007)
21. Rogers, E.M., Bhowmik D.K.: Homophily-heterophily: Relational concepts for communication research.: Public. Opin. Quart. **34**(4), 523–538 (1970)
22. Vapnik, V.: The Nature of Statistical Learning. Data Min. Knowl. Discov. **6**, 1–47
23. Warner, Stanley L.: Randomized response: A survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. **60**(309), 63–69 (1965)
24. Williams, D., Ducheneaut, N., Xiong, L., Zhang, Y., Yee, N., Nickell, E.: From tree house to barracks the social life of guilds in world of Warcraft. Games Cult. **1**(4), 338–361 (2006)
25. Williams, D., Yee N., Caplan S.: Who plays, how much, and why? A behavioral player census of a virtual world. J. Comp. Mediat. Commun. **13**(4) 993–1018 (2008)
26. Williams, D.: The mapping principle, and a research framework for virtual worlds. Commun. Theory **20**(4), 451–470 (2010)
27. Yee, N., Bailenson, J.: The proteus effect: The effect of transformed self-representation on behavior. Hum. Commun. Res. **33**, 271–290 (2007)
28. Zastrow, C., Kirst-Ashman K.K.: Understanding human behavior and the social environment. Cengage Learning, Belmont (2007)