В поисках утраченных профилей:

достоверность данных «ВКонтакте» и их значение для исследований образования

И. Б. Смирнов, Е. В. Сивак, Я. Я. Козьмина

Статья поступила в редакцию в марте 2016 г.

Смирнов Иван Борисович
стажер-исследователь Института образования Национального исследовательского университета «Высшая школа экономики». E-mail: ibsmirnov@hse.ru
Сивак Елизавета Викторовна
научный сотрудник Института образования Национального исследовательского университета «Высшая школа экономики». E-mail: esivak@hse.ru
Козьмина Яна Яковлевна
младший научный сотрудник Института образования Национального исследовательского университета «Высшая школа экономики». E-mail: ikozmina@hse.ru

Адрес: 101000, Москва, ул. Мясницкая, 20.

Аннотация. Потенциал социальной сети «ВКонтакте» как источника информации начинает использоваться в исследованиях образования, однако о качестве данных, которые можно получить с помощью этой социальной сети, и о смещениях выборки ее пользователей относительно общей популяции учащихся до сих пор известно недостаточно. Исследуется достоверность данных «ВКонтакте» на примере одной школы (766 учащихся) и одного университета (15 757 учащихся). Описывается процедура сопоставления профилей на «ВКонтакте» с реальными школьниками и студентами. Непосредственное сопоставление позволяет обнаружить около 18% учащихся. Предлагаемый в статье метод дает возможность увеличить этот показатель до 88% для школьников и 93% для студентов. Оцениваются различия между теми учащимися, которые были найдены на «ВКонтакте», и теми, которые не были найдены, по ряду существенных характеристик, таких как успеваемость, пол и возраст. Анализируется соответствие социальных связей, восстановленных по данным «ВКонтакте», реальной социальной структуре образовательного учреждения. Показано, что в виртуальном пространстве воспроизводится разделение университета на кампусы и на образовательные программы, а школы — на корпуса и классы. Полученные результаты вносят вклад в решение важных для данной научной области вопросов: насколько сведения из социальной сети соответствуют реальности, как можно повысить их точность и как они могут использоваться в исследованиях образования.

Ключевые слова: социальные сети, «ВКонтакте», анализ социальных сетей, достоверность данных, сети дружбы, академическая успеваемость, школы, вузы.

DOI: 10.17323/1814-9545-2016-4-106-122

Авторы выражают признательность анонимному рецензенту журнала «Вопросы образования» за ценные замечания.

Социальные сети стали неотъемлемой частью повседневной жизни миллионов людей, которые используют их для общения с друзьями, обмена идеями, поиска работы, организации мероприятий и многого другого [Boyd, Ellison, 2008]. Крупнейшая социальная сеть Facebook была основана всего 12 лет назад, а сегодня ей пользуются полтора миллиарда человек[1]. Неудивительно, что внимание исследователей привлекает вопрос о влиянии социальных сетей на различные области жизни, включая образование [Hew, 2011; Aydin, 2012; Wilson, Gosling, Graham, 2012; Tess, 2013; Королева, 2015].

Особое внимание к социальным сетям связано еще и с тем, что они произвели революцию в доступности демографических и социальных данных [Boyd, Ellison, 2008]. Даже наиболее масштабные исследования в области образования редко вовлекают больше нескольких десятков тысяч человек, большинство же ограничивается гораздо меньшими выборками. Крупнейшее международное исследование школьников PISA в 2012 г. охватило 510 тыс. школьников из 62 стран мира [OECD, 2014], а А. Крамер с соавторами [Kramer, Guillory, Hancock, 2014] опубликовал результаты эксперимента, проведенного в Facebook, участниками которого стали 700 тыс. человек. Наиболее же масштабный эксперимент на платформе Facebook включал 61 млн человек [Bond et al., 2012].

Социальные сети не только позволяют проводить исследования в недоступных ранее масштабах, они также дают возможность отвечать на новые вопросы. Так, сети дружбы учащихся и эффекты сообучения традиционно исследуются с помощью опросов [Lomi et al., 2011; Flashman, 2012; Иванюшина, Александров, 2013; Докука, Валеева, Юдкевич, 2015]. Этот метод, однако, не позволяет устанавливать и изучать связи между учащимися разных образовательных учреждений. Эти отношения до недавних пор оставались слепым пятном для исследователей образования, сведения же из социальных сетей позволяют их обнаружить и изучать. Социальные сети также открывают доступ к лонгитюдным данным о социальных связях, позволяя не только получить информацию о текущем состоянии, но и проследить всю историю взаимодействия пользователей [Lazer et al., 2009].

В фокусе международных исследований традиционно находится Facebook как наиболее популярная социальная сеть в мире. В США 71% населения, имеющего доступ к Интернету, пользуются Facebook [Duggan et al., 2015]. Для отдельных категорий этот показатель значительно выше. Так, в некоторых университетах 96% студентов используют Facebook [Martin, 2009].

[1] Facebook (2015) Statistics. Facebook, Palo Alto, CA. http://newsroom.fb.com/company-info

Изучается влияние интенсивности использования этой социальной сети на интеграцию студентов в социальную жизнь университета [Madge et al., 2009], социальный капитал студентов [Ellison, Steinfield, Lampe, 2007; Steinfield, Ellison, Lampe, 2008] и их психологическое благополучие [Steinfield, Ellison, Lampe, 2008]. В России аналогом Facebook является социальная сеть «ВКонтакте». Возможности этой сети как источника данных также начинают привлекать внимание исследователей. В частности, они выясняют, как время, которое студенты проводят на «ВКонтакте» перед экзаменами, влияет на их оценки [Krasilnikov, Semenova, 2014], как формируется сеть дружбы студентов [Dokuka, Valeeva, Yudkevich, 2015]. Показано, как данные «ВКонтакте» можно использовать для анализа образовательной мобильности [Alexandrov, Karepin, Musabirov, 2016].

Однако примеров использования данных социальной сети в исследованиях образования пока немного. Их применение затрудняется тем, что до сих пор нет достаточной информации о степени достоверности данных «ВКонтакте» и возможных смещениях выборки пользователей сайта. Так, например, школа № 1 Санкт-Петербурга печально знаменита тем, что — если верить данным «ВКонтакте» — в 2019 г. должна выпустить 3000 школьников. Трудности возникают и при попытке прямого сопоставления списка учащихся с профилями в социальной сети. Школьники и студенты не всегда указывают в профиле свое образовательное учреждение и часто используют альтернативные формы своего имени.

Наша статья посвящена исследованию достоверности данных «ВКонтакте» на примере одной московской школы и одного университета. На первом этапе были получены списки школьников, содержащие информацию об их среднем балле, половой принадлежности, классе и корпусе школы, в котором они учатся, и списки студентов с информацией об успеваемости, курсе и образовательной программе. Затем был произведен поиск профилей учащихся на «ВКонтакте». Прямое сопоставление (точное совпадение имени и фамилии и указание учебного заведения в профиле) позволило обнаружить лишь около 18% учащихся. Использование информации о дружеских связях, а также словаря, включающего разные формы одного имени, дало возможность увеличить этот показатель до 88% для школьников и до 93% для студентов. Было произведено сравнение групп учащихся, найденных разными методами, а также учащихся, которые не были обнаружены на «ВКонтакте». Дополнительно была загружена информация о дружеских связях и проведено сравнение восстановленной по ним структуры образовательного учреждения с реальной.

Нам удалось продемонстрировать возможность извлечения из «ВКонтакте» данных, характеризующихся высокой степенью достоверности, а также соответствие структуры социальных связей, восстановленных по этим данным, структуре образователь-

ного учреждения, включая разделение школы на корпуса и классы и университета — на кампусы и образовательные программы. Насколько нам известно, это первое исследование такого рода и масштаба на данных «ВКонтакте». Полученные результаты позволят исследователям образования с большей эффективностью использовать потенциал социальной сети.

1. Программное обеспечение и процедура поиска данных в сети «ВКонтакте»

Пользователи Интернета, регистрируясь в социальной сети «ВКонтакте», принимают условия пользовательского соглашения, согласно которому они «осознают, что информация на сайте, размещаемая пользователем о себе, может становиться доступной для других пользователей сайта и пользователей Интернета, может быть скопирована и распространена такими пользователями»[2]. «ВКонтакте», в свою очередь, предоставляет API (публичный интерфейс приложения), который позволяет автоматически выполнять поисковые запросы и получать информацию о пользователях, если она не была скрыта настройками приватности.

Разработанное нами программное обеспечение (программа) выполняет запросы к API «ВКонтакте» и получает список всех пользователей, указавших, что они учатся в заданном учебном заведении, и соответствующих определенным возрастным ограничениям. Затем производится сопоставление найденных профилей со списком учащихся, предоставленным образовательным учреждением, по имени и фамилии. Однако прямое сопоставление позволяет обнаружить в социальной сети лишь незначительную часть учащихся. Чтобы извлечь из нее больше информации, мы применили два дополнительных приема.

Во-первых, мы создали словарь альтернативных форм имени. Если программа обнаруживала, что фамилия, указанная в профиле, содержит латинские буквы, она предлагала оператору перевести ее. Таким образом удалось выявить пользователей, указавших свою фамилию латиницей, например «Nabokov» вместо «Набоков». Если программа находила одну и ту же фамилию в списке учащихся и в списке пользователей, она уточняла у оператора, совпадают ли имена. В результате были установлены пользователи, использующие сокращенную форму имени, например «Вова Набоков» вместо «Владимир Набоков». Все переводы и отмеченные совпадения (или несовпадения) имен сохранялись в специальный словарь, и повторно оператору не требовалось отвечать на один и тот же вопрос.

Во-вторых, программа осуществляла поиск не только по пользователям, указавшим в профиле заданное учебное заведение,

[2] ВКонтакте (2016) Правила защиты информации о пользователях сайта VK.com. https://vk.com/privacy

но и по тем пользователям, у которых много друзей из этого учебного заведения. Этот прием, традиционный для анализа социальных сетей, используется, например, в [Mislove et al., 2010].

Для того чтобы обеспечить сохранность личных данных школьников, мы разработали специальную версию программы, которая запускается локально на школьном компьютере и после выполнения процедуры сопоставления удаляет все имена, фамилии и идентификаторы «ВКонтакте». Только полностью обезличенные данные передаются для дальнейшего исследования. Информация о студентах университета (списки студентов, обучающихся на разных образовательных программах, сведения об их успеваемости) была получена из открытых источников (с сайта университета). После проведения процедуры сопоставления имена студентов и идентификаторы были удалены, и в дальнейшем использовался только обезличенный набор данных.

По итогам процедуры сопоставления можно выделить несколько групп учащихся: те, которые не были обнаружены на «ВКонтакте»; те, которые были выявлены непосредственным сопоставлением; те, которые были установлены с помощью предложенного нами метода. Мы сравниваем эти группы по численности, а также по полу, возрасту и успеваемости входящих в них студентов. Для вычисления p-значения используются критерий χ-квадрат и критерий Стьюдента.

Мы также построили сети дружбы учащихся и сопоставили их со структурой образовательных учреждений. Мы ожидаем, что ученики из одной параллели, сокурсники и обучающиеся на одной образовательной программе окажутся тесно связаны между собой. Чтобы выразить эффект от такого разбиения на группы количественно, мы вычисляем модулярность Q. Эта величина равна доле дружеских связей, соединяющих учеников из одной группы (одной параллели, одной образовательной программы и т. п.), минус ожидаемое количество таких связей в том случае, если бы они распределялись случайно. $Q = 0$ означает отсутствие предпочтений образовывать связи внутри своей группы. Чем ближе Q к 1 (максимальное значение), тем сильнее выражено разбиение на группы. На практике Q принимает значения от 0,3 до 0,7, более высокие значения встречаются редко [Newman, Girvan, 2004].

2. Достоверность данных

2.1. Школа

С использованием API «ВКонтакте» мы обнаружили 908 пользователей, указавших в профиле, что им не более 18 лет и что они учатся в исследуемой школе. При этом согласно списку в 5–11-х классах школы учатся 766 учеников. Таким образом, как минимум часть пользователей предоставила о себе ложную информацию.

Эффективным критерием идентификации настоящих профилей школьников может послужить число друзей на «ВКонтакте», указавших в профиле школу с тем же номером. Так, среди

Таблица 1. **Доля учеников, чьи профили были обнаружены на «ВКонтакте» с использованием предложенных методов** (%)

		Словарь альтернативных форм имени: Нет	Словарь альтернативных форм имени: Да
Список друзей	Нет	18	27
	Да	57	88

Таблица 2. **Сравнение долей найденных на «ВКонтакте», не указавших школу и использующих альтернативную форму имени в группах школьников, различающихся по возрасту (классу)**

	Класс						
Доля учащихся (%)	5-й	6-й	7-й	8-й	9-й	10-й	11-й
Всех найденных	85	89	88	90	88	91	85
Не указавших школу	64	72	69	74	70	58	72
Использовавших альтернативную форму имени	39	36	29	33	33	31	38

458 пользователей, у которых нет ни одного друга из той же школы, только четверо, т. е. меньше 1%, являются реальными учениками школы. Среди тех, кто входит в список занимающих первые сто мест по числу друзей в школе, как минимум 83% учатся в данной школе (табл. 1).

Итоговый охват сопоставим с полученным в исследовании, в котором анализировались профили американских студентов на Facebook, — там, согласно выложенным в открытый доступ данным, охват во второй волне составил 84,6% [Lewis et al., 2008].

В табл. 2 представлено сравнение групп учащихся, различающихся по возрасту (классу). Примерно одинаковые доли учащихся были обнаружены в социальной сети для всех параллелей. Использование альтернативной формы имени и указание школы в профиле также не меняется от параллели к параллели. Ни одно из различий, указанных в таблице, не достигает уровня значимости. *p*-значения, вычисленные по критерию χ-квадрат, больше 0,5.

Точно так же не наблюдается различий по половому составу и успеваемости между группами школьников, найденными на «ВКонтакте», не найденными на «ВКонтакте»», не указавшими

Таблица 3. **Сравнение групп учащихся, различающихся способом представления данных о себе на «ВКонтакте», по половому составу и успеваемости**

	Девочки (%)	Средний балл
Найденные на «ВКонтакте»	46	3,80
Не найденные на «ВКонтакте»	48	3,79
Не указавшие школу	48	3,77
Использующие альтернативную форму имени	50	3,79

Таблица 4. **Сравнение долей найденных на «ВКонтакте» и использующих альтернативную форму имени в группах студентов, различающихся по возрасту (курсу)**

Доля студентов (%)	Курс			
	1-й	2-й	3-й	4-й
Найденных	92	94	94	93
Использовавших альтернативную форму имени	30	32	32	34

Таблица 5. **Сравнение групп студентов, различающихся способом представления данных о себе на «ВКонтакте», по половому составу и успеваемости**

	Девушки (%)	Средний балл
Найденные на «ВКонтакте»	59	7,34
Не найденные на «ВКонтакте»	58	7,13
Использовавшие альтернативную форму имени	71	7,37

школу и использующими альтернативную форму имени по полу или успеваемости (табл. 3), p-значения больше 0,5.

2.2. Университет

Аналогичные результаты были получены и для студентов университета. Из 15757 студентов 93% были обнаружены на «ВКонтакте». В зависимости от образовательной программы этот показатель варьирует от 75 до 100%.

Между студентами, найденными и не найденными на «ВКонтакте», а также между использующими и не использующими альтернативную форму имени, нет различий по возрасту, однако ненайденные на «ВКонтакте» студенты в среднем учатся несколько хуже (p-значение $< 10^{-8}$), а девушки чаще используют альтернативные формы имени, чем юноши (p-значение $< 10^{-11}$).

Рис. 1. **Сеть дружбы на «ВКонтакте» воспроизводит разделение школы на классы. Ученики из одного класса в основном дружат между собой. Чем больше разница в возрасте между учениками, тем меньше вероятность дружбы между ними**

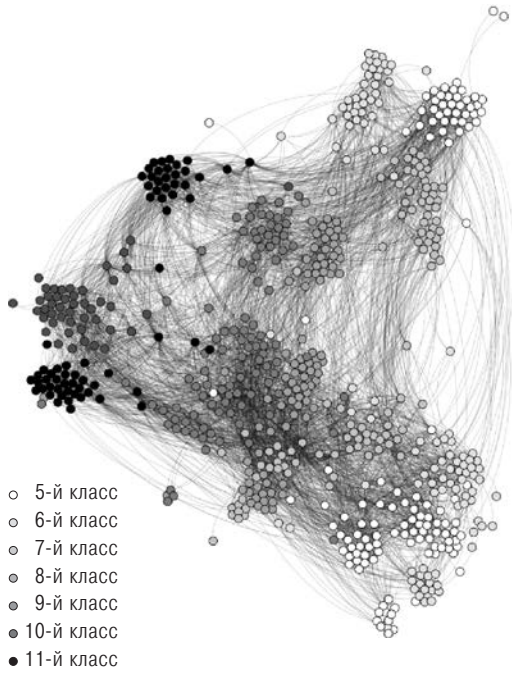
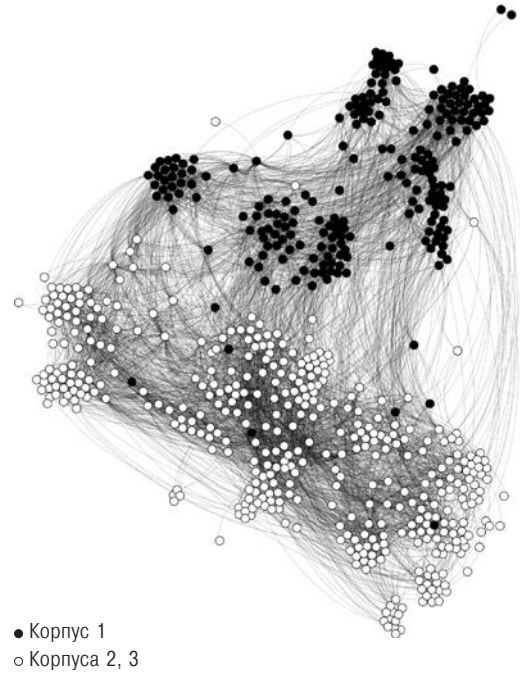


Рис. 2. **Сеть дружбы на «ВКонтакте» воспроизводит разбиение школы на несколько корпусов**



Использование альтернативных форм имени различается у школьников и студентов. Так, 27% всех альтернативных форм имени, используемых студентами, — это имена, набранные латиницей, у школьников же доля таких имен — только 8%.

3. Структура сети дружбы

3.1. Школа

Для всех школьников, обнаруженных на «ВКонтакте», мы построили сеть их дружбы (рис. 1). Для визуализации сети использовался алгоритм Force Atlas 2 и программное обеспечение Gephi [Jacomy et al., 2014]. Алгоритм располагает узлы сети тем ближе друг к другу, чем теснее они связаны между собой. Полученная структура сети соответствует разделению на параллели классов, модулярность $Q = 0{,}47$, при этом дистанция зависит от разницы в возрасте: на наибольшем удалении друг от друга находятся младшие и старшие классы. Дополнительно сеть дружбы разбивается на два больших кластера, соответствующих разным корпусам недавно объединенных школ, $Q = 0{,}35$ (рис. 2).

Рис. 3. **Сеть дружбы на «ВКонтакте» воспроизводит разделение образовательной программы на курсы. Чем больше разница в возрасте между студентами, тем меньше вероятность дружбы между ними**

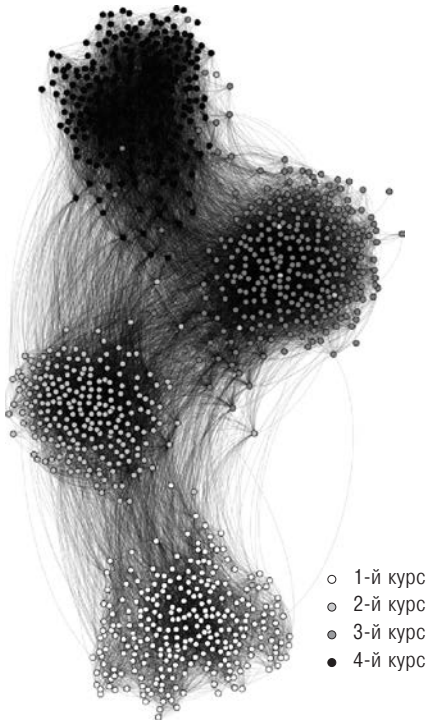
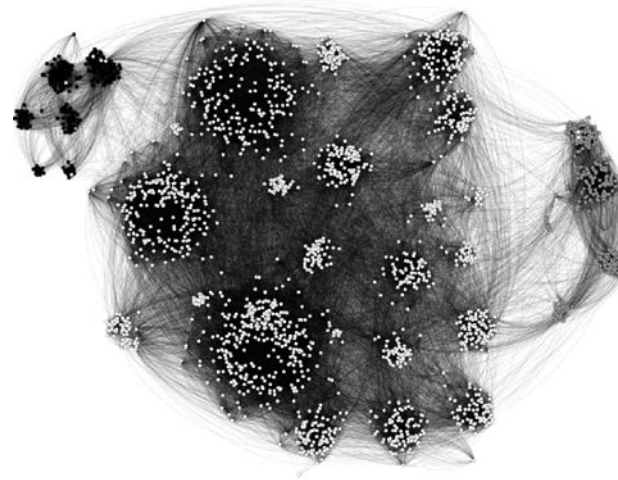


Рис. 4. **Сеть дружбы на «ВКонтакте» воспроизводит разделение университета на кампусы, находящиеся в разных городах, и на образовательные программы. На рисунке представлена сеть дружбы четверокурсников. Видимые кластеры внутри кампусов соответствуют образовательным программам**



3.2. Университет

Сеть дружбы студентов на «ВКонтакте» воспроизводит разделение на курсы, $Q = 0{,}58$ (рис. 3), кампусы, $Q = 0{,}32$, и образовательные программы, $Q = 0{,}68$ (рис. 4).

4. Перспективы использования данных «ВКонтакте» в исследованиях образования

«ВКонтакте» как источник данных обладает большим потенциалом для исследований образования. Однако использование этих данных сопряжено с рядом методологических трудностей. Результаты нашей работы позволяют дать конкретные рекомендации по их преодолению.

Из списка пользователей, которые представляются учащимися заданного учебного заведения, целесообразно исключать тех, у кого нет на «ВКонтакте» друзей из этого же учебного заведения. Только 1% таких пользователей действительно учатся в нем.

При сопоставлении списка учащихся со списком пользователей «ВКонтакте» нужно учитывать альтернативные формы имени, так как их используют 35% учащихся. Эффективным средством пополнения списка учащихся — пользователей социальной сети является дополнительный поиск среди друзей обнаруженных пользователей, так как 69% учащихся не указывают в профиле свое образовательное учреждение.

Особое внимание при использовании данных из социальных сетей следует уделить потенциальным смещениям выборки. Например, можно ожидать, что ученики младших классов будут менее представлены на «ВКонтакте», чем ученики старших классов, что наименее успевающие ученики будут чаще указывать в профиле не соответствующие действительности сведения и т. п. Однако в данном исследовании значимых различий по половому составу, возрасту и успеваемости между группами найденных и не найденных на «ВКонтакте» среди школьников 5–11-х классов и студентов обнаружено не было. Исключение составляет чуть меньший средний балл у студентов, не найденных на «ВКонтакте», по сравнению с найденными в социальной сети и более частое использование альтернативных форм имени девушками по сравнению с юношами.

Итоговый охват в 88% школьников и 93% студентов свидетельствует о том, что социальной сетью пользуются практически все учащиеся. Представляет интерес воспроизведение наших результатов на большей выборке и в особенности сравнение разных регионов и населенных пунктов.

Результаты нашего исследования подтверждают и ценность информации о дружеских связях на «ВКонтакте». Мы показали, что структура этих связей соответствует социальной структуре реального учебного заведения: она воспроизводит не только распределение учащихся на классы, курсы и образовательные программы, но и пространственную структуру учебного заведения, такую как разделение школы на несколько корпусов.

Социальные сети позволяют по-новому взглянуть на традиционные для исследований образования темы. С конца 1970-х годов набирает силу традиция изучения социального и культурного капитала [Bourdieu, 1986; Coleman, 1988; Putnam, 2001], эти конструкты доказали свою значимость и в исследованиях образования [DiMaggio, 1982; Goddard, 2003; Lareau, Weininger, 2003]. Особое внимание уделяется при этом воспроизводству неравенства [Bourdieu, Passeron, 1990; Stanton-Salazar, Dornbusch, 1995]. Сегодня появляется уникальная возможность проверить социологические теории на новых масштабных эмпирических данных.

Информацию о культурном капитале школьников можно реконструировать через указанные в профиле интересы, через подписки на группы и страницы «ВКонтакте», характеризующие вкусы и культурные предпочтения школьников [Liu, 2007; Lewis et

al., 2012]. Что касается социального капитала, то данные из социальных сетей позволяют отслеживать как слабые связи (дружба на «ВКонтакте»), так и сильные (комментирование записей друг друга, отметка «Мне нравится» и т. п.). При этом по своим масштабам и детальности такие данные значительно превосходят результаты социометрических исследований, которые чаще всего не выходят за рамки контактов внутри одного класса, игнорируя межвозрастные и межшкольные связи.

Социальные сети позволяют исследовать связь культурного и социального капитала с образовательными достижениями как на уровне школ, так и на уровне отдельных учеников. При этом становится возможным не только зафиксировать наличие географической и социальной сегрегации и ее отражение в виртуальном пространстве, но и изучать механизмы воспроизводства неравенства: влияние школьников друг на друга (эффекты сообучения, влияние друзей на установки школьников и т. п.), влияние культурного и социального капитала на выбор образовательной траектории (смена школы, переход из школы в вуз).

Использование данных из социальных сетей не только открывает новые возможности перед исследователями образования, но и ставит перед ними новые этические вопросы. В социальных сетях доступность информации о пользователе больше не зависит только от того, какую информацию он сам решил разместить. Например, можно с большой степенью точности восстановить информацию об университете, годе выпуска и специальности [Mislove et al., 2010], сексуальной ориентации [Bhattasali, Maiti, 2015], романтическом партнере [Backstrom, Kleinberg, 2014] или политических убеждениях [Bakshy, Messing, Adamic, 2015] пользователя. В своей работе мы показываем, что даже «наивные» средства позволяют определить номер школы тех учащихся, которые решили его не указывать на «ВКонтакте». Применение продвинутых алгоритмов машинного обучения позволит это сделать еще эффективнее. Данные из социальных сетей зачастую требуется объединить с дополнительными сведениями, полученными из открытых источников или от учебных учреждений. Процедура такого сопоставления требует особого внимания к обезличиванию данных, гарантирующему сохранность личной информации.

Литература

1. Иванюшина В. А., Александров Д. А. Антишкольная культура и социальные сети школьников // Вопросы образования. 2013. № 2. С. 233–252.
2. Королева Д. О. Использование социальных сетей в образовании и социализации подростка: аналитический обзор эмпирических исследований (международный опыт)// Психологическая наука и образование. 2015. Т. 20. № 1. С. 28–37.
3. Alexandrov D., Karepin V., Musabirov I. (2016) Educational Migration from Russia to China: Social Network Data/ Proceedings of the 8th ACM Con-

ference on Web Science, May 22 to May 25, 2016, Hannover, Germany. P. 309–311.
4. Aydin S. (2012) A Review of Research on Facebook as an Educational Environment // Educational Technology Research and Development. Vol. 60. No 6. P. 1093–1106.
5. Backstrom L., Kleinberg J. (2014) Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook/ Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, February 15–19, 2014, Baltimore, Maryland, USA. P. 831–841.
6. Bakshy E., Messing S., Adamic L. A. (2015) Exposure to Ideologically Diverse News and Opinion on Facebook // Science. Vol. 348. No 6239. P. 1130–1132.
7. Bhattasali N., Maiti E. (2015) Machine «Gaydar»: Using Facebook Profiles to Predict Sexual Orientation. http://cs229.stanford.edu/proj2015/019_report.pdf
8. Bond R. M., Fariss C. J., Jones J. J., Kramer A. D., Marlow C., Settle J. E., Fowler J. H. (2012) A 61-Million-Person Experiment in Social Influence and Political Mobilization // Nature. Vol. 489. No 7415. P. 295–298.
9. Bourdieu P. (1986) The Forms of Capital // Cultural Theory: An Anthology. P. 81–93.
10. Bourdieu P., Passeron J. C. (1990) Reproduction in Education, Society and Culture (Theory, Culture & Society). London: Sage Publications.
11. Boyd D. M., Ellison N. B. (2008) Social Network Sites: Definition, History, and Scholarship // Journal of Computer-Mediated Communication.Vol. 13. No 1. P. 210–230.
12. Christakis N. A., Fowler J. H. (2013) Social Contagion Theory: Examining Dynamic Social Networks and Human Behavior // Statistics in Medicine. Vol. 32. No 4. P. 556–577.
13. Coleman J. S. (1988) Social Capital in the Creation of Human Capital // American Journal of Sociology. Vol. 94. No 1. P. 95–120.
14. DiMaggio P. (1982) Cultural Capital and School Success: The Impact of Status Culture Participation on the Grades of US High School Students // American Sociological Review. Vol. 47. No 2. P. 189–201.
15. Dokuka S., Valeeva D., Yudkevich M. (2015) Formation and Evolution Mechanisms in Online Network of Students: The Vkontakte Case // M. Y. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, V. G. Labunets (eds) Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science. Vol. 542. P. 263–274.
16. Duggan M., Ellison N. B., Lampe C., Lenhart A., Madden M. (2015) Social Media Update 2014. http://www.pewinternet.org/2015/01/09/social-media-update-2014/
17. Ellison N. B., Steinfield C., Lampe C. (2007) The Benefits of Facebook «Friends:» Social Capital and College Students' Use of Online Social Network Sites // Journal of Computer-Mediated Communication. Vol. 12. No 4. P. 1143–1168.
18. Flashman J. (2012) Academic Achievement and Its Impact on Friend Dynamics // Sociology of Education. Vol. 85. No 1. P. 61–80.
19. Goddard R. D. (2003) Relational Networks, Social Trust, and Norms: A Social Capital Perspective on Students' Chances of Academic Success // Educational Evaluation and Policy Analysis. Vol. 25. No 1. P. 59–74.
20. Hew K. F. (2011) Students' and Teachers' Use of Facebook // Computers in Human Behavior. Vol. 27. No 2. P. 662–676.

21. Jacomy M., Venturini T., Heymann S., Bastian M. (2014) Force Atlas 2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software // PLoS ONE. Vol. 9. No 6. e98679.
22. Kramer A. D., Guillory J. E., Hancock J. T. (2014) Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks // Proceedings of the National Academy of Sciences. Vol. 111. No 24. P. 8788–8790.
23. Krasilnikov A., Semenova M. (2014) Do Social Networks Help to Improve Student Academic Performance? The Case of Vk.com and Russian Students // Economics Bulletin. Vol. 34. No 2. P. 718–733.
24. Lazer D., Pentland A. S., Adamic L., Aral S., Barabasi A. L., Brewer D., Jebara T. (2009) Life in the Network: The Coming Age of Computational Social Science // Science. Vol. 323. No 5915. P. 721–723.
25. Lareau A., Weininger E. B. (2003) Cultural Capital in Educational Research: A Critical Assessment // Theory and Society. Vol. 32. No 5–6. P. 567–606.
26. Lewis K., Gonzalez M., Kaufman J. (2012) Social Selection and Peer Influence in an Online Social Network // Proceedings of the National Academy of Sciences. Vol. 109. No 1. P. 68–72.
27. Lewis K., Kaufman J., Gonzalez M., Wimmer A., Christakis N. (2008) Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.com // Social Networks. Vol. 30. No 4. P. 330–342.
28. Liu H. (2007) Social Network Profiles as Taste Performances // Journal of Computer-Mediated Communication. Vol. 13. No 1. P. 252–275.
29. Lomi A., Snijders T. A., Steglich C. E., Torló V. J. (2011) Why Are Some More Peer than Others? Evidence from a Longitudinal Study of Social Networks and Individual Academic Performance // Social Science Research. Vol. 40. No 6. P. 1506–1520.
30. Madge C., Meek J., Wellens J., Hooley T. (2009) Facebook, Social Integration and Informal Learning at University:'It Is More for Socialising and Talking to Friends about Work than for Actually Doing Work' // Learning, Media and Technology. Vol. 34. No 2. P. 141–155.
31. Marginson S. (2014) University Rankings and Social Science // European Journal of Education. Vol. 49. No 1. P. 45–59.
32. Martin C. (2009) Social Networking Usage and Grades among College Students.http://www.pdfpedia.com/download/15925/social-networking-usage-and-grades-among-college-students-pdf.html
33. Mislove A., Viswanath B., Gummadi K. P., Druschel P. (2010) You Are Who You Know: Inferring User Profiles in Online Social Networks. Proceedings of the Third ACM International Conference on Web Search and Data Mining, February 3–5, 2010, New York City, USA. P. 251–260.
34. Newman M. E., Girvan M. (2004) Finding and Evaluating Community Structure in Networks // Physical Review E. Vol. 69. No 2. 026113.
35. OECD (2014) PISA 2012 Technical Report. OECD: Paris.
36. Putnam R. (2001) Social Capital: Measurement and Consequences // Canadian Journal of Policy Research. Vol. 2. No 1. P. 41–51.
37. Stanton-Salazar R.D., Dornbusch S. M. (1995) Social Capital and the Reproduction of Inequality: Information Networks among Mexican-Origin High School Students // Sociology of Education. Vol. 68. No 2. P. 116–135.
38. Steinfield C., Ellison N. B., Lampe C. (2008) Social Capital, Self-Esteem, and Use of Online Social Network Sites: A Longitudinal Analysis // Journal of Applied Developmental Psychology. Vol. 29. No 6. P. 434–445.
39. Tess P. A. (2013) The Role of Social Media in Higher Education Classes (Real and Virtual) — A Literature Review // Computers in Human Behavior. Vol. 29. No 5. P. A60–A68.

40. Wilson R. E., Gosling S. D., Graham L. T. (2012) A Review of Facebook Research in the Social Sciences // Perspectives on Psychological Science. Vol. 7. No 3. P. 203–220.
41. Whitley B., Keith-Spiegel P. (2002) Academic Dishonesty: An Educators Guide. New Jersey: Lawrence Erlbaum Associates.

In Search of Lost Profiles: The Reliability of VKontakte Data and Its Importance for Educational Research

Authors

Ivan Smirnov
ResearchAssistant, Institute of Education, National Research University Higher School of Economics. E-mail: ibsmirnov@hse.ru

Elizaveta Sivak
Research Fellow, Institute of Education, National Research University Higher School of Economics. E-mail: esivak@hse.ru

Yana Kozmina
Junior Research Fellow, Institute of Education, National Research University Higher School of Economics. E-mail: ikozmina@hse.ru

Address: 20 Myasnitskaya str., 101000 Moscow, Russian Federation.

Abstract

The potential of VKontakte as a data source is now acknowledged in educational research, but little is known about the reliability of data obtained from this social network and about its sampling bias. Our article investigates the reliability of VK data, using the examples of a secondary school (766 students) and a university (15,757 students). We describe the procedure of matching VK profiles to real students. A direct comparison permitted us to identify profiles of around 18% of students. A special technique introduced in the article increased this number up to 88% for school students and up to 93% for university students. We compare age, gender and GPA of identified students and those whomwe did not find on VK. We also compare the structure of social relationships, retrieved from VK data, to the expected structure of students' social ties. We found that the structure of 'virtual' social relationships reproduces both the socio-demographic division of students into classes or majors and the spatial division into different school buildings or university campuses. To our knowledge, it is the first study of this kind and scale based on VK data. It contributes to the understanding of how reliable data from this SNS is, how its accuracy can be improved, and how it can be used in educational research.

Keywords

social network analysis, social network sites, VK, datareliability, friendship networks, academic achievement, school.

References

Alexandrov D., Karepin V., Musabirov I. (2016) Educational Migration from Russia to China: Social Network Data. Proceedings of the *8th ACM Conference on Web Science, May 22 to May 25, 2016, Hannover, Germany*, pp. 309–311.

Aydin S. (2012) A Review of Research on Facebook as an Educational Environment. *Educational Technology Research and Development*, vol. 60, no 6, pp. 1093–1106.

Backstrom L., Kleinberg J. (2014) Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook. Proceedings of the *17th ACM Conference on Computer Supported Cooperative Work & Social Computing, February 15–19, 2014, Baltimore, Maryland, USA*, pp. 831–841.

Bakshy E., Messing S., Adamic L. A. (2015) Exposure to Ideologically Diverse News and Opinion on Facebook. *Science*, vol. 348, no 6239, pp. 1130–1132.

Bhattasali N., Maiti E. (2015) *Machine "Gaydar": Using Facebook Profiles to Predict Sexual Orientation*. Available at: http://cs229.stanford.edu/proj2015/019_report.pdf (accessed 10 October 2016).

Bond R. M., Fariss C. J., Jones J. J., Kramer A. D., Marlow C., Settle J. E., Fowler J. H. (2012) A 61-Million-Person Experiment in Social Influence and Political Mobilization. *Nature*, vol. 489, no 7415, pp. 295–298.

Bourdieu P. (1986) The Forms of Capital. *Cultural Theory: An Anthology*, Malden, MA: Wiley-Blackwell, pp. 81–93.

Bourdieu P., Passeron J. C. (1990) *Reproduction in Education, Society and Culture (Theory, Culture & Society)*. London: Sage Publications.

Boyd D. M., Ellison N. B. (2008) Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, vol. 13, no 1, pp. 210–230.

Christakis N. A., Fowler J. H. (2013) Social Contagion Theory: Examining Dynamic Social Networks and Human Behavior. *Statistics in Medicine*, vol. 32, no 4, pp. 556–577.

Coleman J. S. (1988) Social Capital in the Creation of Human Capital. *American Journal of Sociology*, vol. 94, no 1, pp. 95–120.

DiMaggio P. (1982) Cultural Capital and School Success: The Impact of Status Culture Participation on the Grades of US High School Students. *American Sociological Review*, vol. 47, no 2, pp. 189–201.

Dokuka S., Valeeva D., Yudkevich M. (2015) Koevolyutsiya sotsialnykh setey i akademicheskikh dostizheniy studentov [Co-Evolution of Social Networks and Student Performance]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 44–65.

Dokuka S., Valeeva D., Yudkevich M. (2015) Formation and Evolution Mechanisms in Online Network of Students: The VKontakte Case. *Analysis of Images, Social Networks and Texts* (eds M.Y. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, V. G. Labunets), pp. 263–274.

Duggan M., Ellison N. B., Lampe C., Lenhart A., Madden M. (2015) *Social Media Update 2014.* Available at: http://www.pewinternet.org/2015/01/09/social-media-update-2014/ (accessed 10 October 2016).

Ellison N. B., Steinfield C., Lampe C. (2007) The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, vol. 12, no 4, pp. 1143–1168.

Flashman J. (2012) Academic Achievement and Its Impact on Friend Dynamics.*Sociology of Education*, vol. 85, no 1, pp. 61–80.

Goddard R. D. (2003) Relational Networks, Social Trust, and Norms: A Social Capital Perspective on Students' Chances of Academic Success. *Educational Evaluation and Policy Analysis*, vol. 25, no 1, pp. 59–74.

Hew K. F. (2011) Students' and Teachers' Use of Facebook. *Computers in Human Behavior,* vol. 27, no 2, pp. 662–676.

Ivaniushina V., Alexandrov D. (2013) Antishkolnaya kultura i sotsialnye seti shkolnikov [Anti-School Culture and Social Networks in Schools]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 2, pp. 233–251.

Jacomy M., Venturini T., Heymann S., Bastian M. (2014) Force Atlas 2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, vol. 9, no 6, e98679.

Koroleva D. (2015) Ispolzovanie sotsialnykh setey v obrazovanii i sotsializatsii podrostka: analiticheskiy obzor empiricheskikh issledovaniy (mezhdunarodny opyt) [Using Social Networks in Education and Socialization of Teenagers: Analytical Review of Empirical Studies (International Experience)]. *Psikhologicheskaya nauka i obrazovanie*, vol. 20, no 1, pp. 28–37.

Kramer A. D., Guillory J. E., Hancock J. T. (2014) Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks. *Proceedings of the National Academy of Sciences*, vol. 111, no 24, pp. 8788–8790.

Krasilnikov A., Semenova M. (2014) Do Social Networks Help to Improve Student Academic Performance? The Case of Vk.com and Russian Students. *Economics Bulletin*, vol. 34, no 2, pp. 718–733.

Lazer D., Pentland A. S., Adamic L., Aral S., Barabasi A. L., Brewer D., Jebara T. (2009) Life in the Network: The Coming Age of Computational Social Science. *Science*, vol. 323, no 5915, pp. 721–723.

Lareau A., Weininger E. B. (2003) Cultural Capital in Educational Research: A Critical Assessment. *Theory and Society*, vol. 32, no 5–6, pp. 567–606.

Lewis K., Gonzalez M., Kaufman J. (2012) Social Selection and Peer Influence in an Online Social Network. *Proceedings of the National Academy of Sciences*, vol. 109, no 1, pp. 68–72.

Lewis K., Kaufman J., Gonzalez M., Wimmer A., Christakis N. (2008) Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.com. S*ocial Networks*, vol. 30, no 4, pp. 330–342.

Liu H. (2007) Social Network Profiles as Taste Performances. *Journal of Computer-Mediated Communication*, vol. 13, no 1, pp. 252–275.

Lomi A., Snijders T. A., Steglich C. E., Torló V. J. (2011) Why Are Some More Peer than Others? Evidence from a Longitudinal Study of Social Networks and Individual Academic Performance. *Social Science Research*, vol. 40, no 6, pp. 1506–1520.

Madge C., Meek J., Wellens J., Hooley T. (2009) Facebook, Social Integration and Informal Learning at University: 'It Is More for Socialising and Talking to Friends about Work than for Actually Doing Work'. *Learning, Media and Technology*, vol. 34, no 2, pp. 141–155.

Marginson S. (2014) University Rankings and Social Science. *European Journal of Education*, vol. 49, no 1, pp. 45–59.

Martin C. (2009) Social Networking Usage and Grades among College Students.Available at: http://www.pdfpedia.com/download/15925/social-networking-usage-and-grades-among-college-students-pdf.html(accessed 10 October 2016).

Mislove A., Viswanath B., Gummadi K. P., Druschel P. (2010) You Are Who You Know: Inferring User Profiles in Online Social Networks. Proceedings of the *Third ACM International Conference on Web Search and Data Mining, February 3–5, 2010, New York City, USA*, pp. 251–260.

Newman M. E., Girvan M. (2004) Finding and Evaluating Community Structure in Networks. *Physical Review E*, vol. 69, no 2, 026113.

OECD (2014) *PISA 2012 Technical Report*. OECD: Paris.

Putnam R. (2001) Social Capital: Measurement and Consequences. *Canadian Journal of Policy Research*, vol. 2, no 1, pp. 41–51.

Stanton-Salazar R.D., Dornbusch S. M. (1995) Social Capital and the Reproduction of Inequality: Information Networks among Mexican-Origin High School Students. *Sociology of Education*, vol. 68, no 2, pp. 116–135.

Steinfield C., Ellison N. B., Lampe C. (2008) Social Capital, Self-Esteem, and Use of Online Social Network Sites: A Longitudinal Analysis. *Journal of Applied Developmental Psychology*, vol. 29, no 6, pp. 434–445.

Tess P. A. (2013) The Role of Social Media in Higher Education Classes (Real and Virtual)—A Literature Review. *Computers in Human Behavior*, vol. 29, no 5, pp. A60–A68.

Wilson R. E., Gosling S. D., Graham L. T. (2012) A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science*, vol. 7, no 3, pp. 203–220.

Whitley B., Keith-Spiegel P. (2002) *Academic Dishonesty: An Educators Guide*. New Jersey: Lawrence Erlbaum Associates.