nanoHUB user behavior: moving from retrospective statistics to actionable behavior analysis

Gerhard Klimeck[1], Gustavo A. Valencia-Zapata[1], Nathan Denny[2], Lynn K. Zentner[1], Michael G. Zentner[1,2]
[1]Network for Computational Nanotechnology
[2]Rosen Center for Advanced Computing
Purdue University, West Lafayette, IN 47907, USA

Abstract

nanoHUB annually serves 17,000+ registered users with over 1 million simulations. In the past, we have used data analytics to demonstrate that nanoHUB *can* be a powerful scientific knowledge sharing platform. We used *retrospective data analytics* to show how simulation tools were used in structured education and how simulation tools were used in novel research. With the use of such *retrospective analytics,* we have made strategic decisions in terms of tool and content developments and justified continued nanoHUB investments by the US National Science Foundation (NSF). As we migrate towards a sustainable nanoHUB we must embrace similar processes pursued by in similar platforms such as Uber or AirBnB: we need to create *actionable data analytics* that can rapidly support user experience and help grow the supply in the two-sided market platform – we need to improve the experience of providers as well as end-users. This paper describes some aspects on how we pursue user behavior analysis inside the virtual worlds of nanotechnology simulation tools. From such user behavior we plan to derive actionable analytics that influence user behaviors as they interact with nanoHUB.

Keywords— nanoHUB; HUBzero; science gateways; user behavior; analytics; cluster; meander; education

Introduction and Background

nanoHUB is a scientific knowledge platform that has enabled over 3,500 researchers and educators to share 500+ research simulation tools and models as well as 6,000+ lectures and tutorials globally through a novel cyberinfrastructure. nanoHUB annually serves 17,000+ registered users with over 1 million simulations in an end-to-end user-oriented scientific computing cloud. Over 1.5 million visitors access the openly available web content items annually. These might be considered impressive summative numbers, but they do not address if the site has any impact or what these users are doing.

Understanding these numbers requires some background on the original intentions and cyberinfrastructure developments around nanoHUB. Fundamental issues raised by peer-reviewers were the perceived ability of a University project to provide a stable, national-level infrastructure, provide support for the offered services, and provide compute cycles for an ever-growing user base.

From the very beginning in 1996 [1], the predecessor to nanoHUB called Purdue Network Computing Hub (PUNCH) was created to enable researchers to share their code without re-writes through novel web interfaces with end-users in education and research. PUNCH was so novel that even the web-server had to be created within the team. By 2004 the standard web-form-interfaces were antiquated and did not inspire the interactive exploration of simulation results with rapid "What If?" questions that users might have. Users had to download their simulation data to manipulate them in a form where they can be truly used. nanoHUB was not an end-to-end usage platform. It became clear that the system had to be revamped to enable the hosting of user-friendly engineering-use inspired interactive applications. Such interactive sessions had to be hosted in a reliable, scalable middleware that was running in production mode, not as a research paper demonstration. 3D dataset exploration had to be supported on remote, dedicated GPUs that deliver the results to end users.

RAPPTURE, the Rapid APPlication infrastrucTURE toolkit [2] enabled researchers, who typically did not have any graphical user interfaces to their codes to describe the input and outputs of their codes in XML and to generate a GUI. New middleware [3] enabled 1,000+ users to be hosted simultaneously on a moderate cluster of about 20 compute nodes. A novel remote GPU-based visualization system [4] supported hundreds of simultaneous sessions. nanoHUB established the first community accounts on TeraGrid and OSG which would execute heavy-lifting nanoHUB simulation jobs completely transparently on behalf of users who had no accounts on these grid platforms [5]. We developed processes [6] to continually test the reliability of these remote grid services to ensure smooth user services. For application support we developed policies and operational infrastructure that enabled tool contributors to support and improve their tools through question & answer forums and through wishlists. As this novel infrastructure emerged in 2005 we observed rapid growth in the simulation user base from the historical numbers of 500 annual users to over 10,000 in a few years. As

questions of technical feasibility were addressed new questions as to actual and potential impact emerged.

Early-on our peer reviewers raised fundamental questions whether such research-based simulation tools could be used by other researchers at all and if these tools could be used in education without specific customizations. The nanoHUB team developed analytics that documented nanoHUB use research through reference and citation searches in the scientific literature. Today we can document over 2,200 papers that cite nanoHUB and we keep track of the used resources and tools, to provide attribution to the published tools. When we showed the first 200 formal citations our peers remained unconvinced that this could be good research. We then began to track secondary citations, which today sum to over 30,000 resulting in an h-index of 82.

Our peers had a similarly strong opinion that research tools could not be used in education. We therefore developed novel clustering algorithms [7] that documented systematic used of simulation tools in formal education settings. Today we can show that over 35,000 students in over 1,800 classes at over 180 institutions have used nanoHUB in formalized education settings. We could also measure the time-to-adoption between tool publication and first-time systematic use in a classroom. The median time was determined to be less than 6 months.

From the analysis of research use and education use we can begin to qualify the attributes of the underlying simulation tools. We found significant use in education and in research for many of the nanoHUB tools. These research and education impact studies are documented in detail in *Nature Nanotechnology* [8].

We used *retrospective* data analytics to show how simulation tools were used in structured education and how simulation tools were used in novel research. We showed that the transition from research tool publication to adoption in the classroom is happening rapidly in typically less than six months and demonstrated through longitudinal data how research tools migrate into education. With the use of these retrospective analytics, we have made strategic decisions in terms of tool and content developments and justified continued investments by NSF into nanoHUB.

As we migrate towards a sustainable nanoHUB we must embrace similar processes pursued by in similar platforms such as Uber or AirBnB: we need to create *actionable data analytics* that can rapidly support user experience and help grow the supply in the two-sided market platform – we need to improve the experience of providers as well as end-users.

II Research Questions

Beyond raw numbers of users and simulations, we have over the years continued to ask ourselves: How do users behave in their virtual world of a simulation tool? More specifically:

- How do they "travel" through the design/exploration world?
- How many individual simulations do they run within one session?
- How many parameters do users change?
- How different do researchers, classroom users, and self-study users behave?
- How different do different classes behave?
- Does different class instruction material / scaffolding make a difference?
- Can we provide feedback to instructors on their classrooms?
- Given certain usage patterns inside the tool:
- Can we improve the tools and provide feedback to the developers?

There are a variety of different requirements that need to be met to address some of these questions in a scalable infrastructure such as:

- Storage/availability of individual simulation runs within user sessions
- A data description language that is shared across different tools
- A large set of simulation runs and participants
- Other user data such as classroom participation, or researcher identification, geolocation, etc.

In the next Sections we describe some of our first results that begin to address some of these questions.

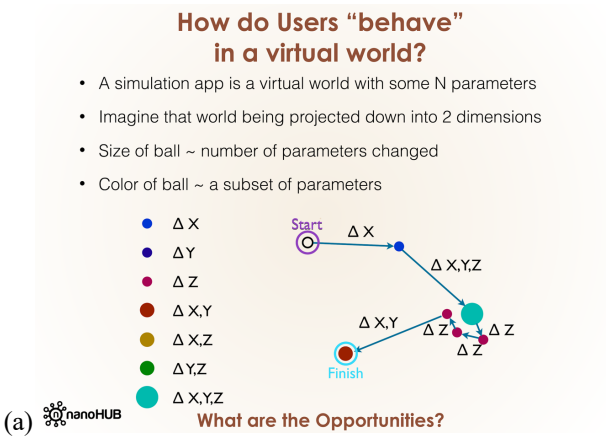
For our initial study presented here we focus on the user behavior for *PN Junction Lab* [9] which is consistently one of the top 10 nanoHUB tools [10] within any year. Despite our codename *pntoy* the tool is powered by an industrial strength semiconductor device modeling tool called PADRE [11]. Instead of learning the complex PADRE input language that involves gridding, geometry, material and environmental specifications, users can easily ask "What if?" questions in a toy-like fashion.

II Searchers and Wildcatters

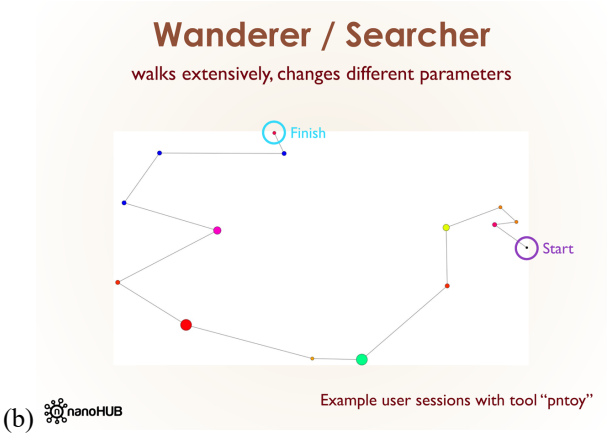
RAPPTURE provides a rather generic description of simulation tool inputs and outputs. Over 90% of the 500+ nanoHUB simulation tools utilize RAPPTURE as their data description language. With existing simulation logs we can now begin to study the user behavior inside simulation tools. Each simulation tool typically consists of 10 to 50 parameters that are exposed to the users. Most of these parameters are freeform numbers such as length, doping, effective mass, dielectric constant, temperature etc. with their specific units, while there is also a significant set of discrete options such as model or geometry choices. Assuming that each parameter might have just 10 reasonable choices, then each tool spans a configurational design space of at least 10^{10} to 10^{50}. The dimensionality of these tools is clearly too large to be intuitively understood.

We developed a visualization methodology [12] to flatten an N-dimensional space into 2 dimensions. Figure 1 shows the conceptual mapping and shows two significantly different user behaviors. A searcher, who moves through the design space in subsequent steps that appear to indicate a method or a goal. A wildcatter who modifies, apparently wildly, the same set of parameters and appears to jump throughout the design space.

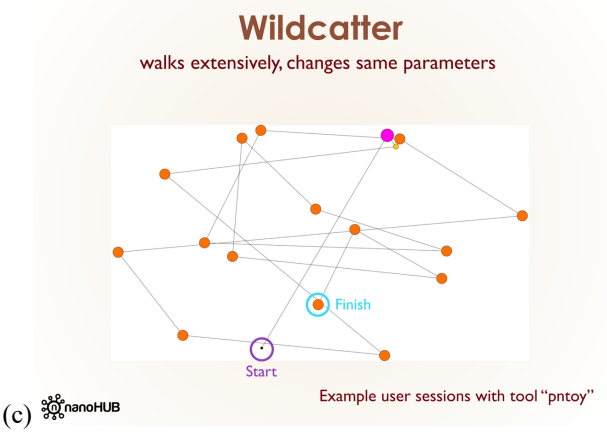
Within the same publication we also documented the development of a "Searchiness" index that assigns a single value to the degree a user behaves like a prototypical wildcatter (Searchiness=0) and prototypical searcher (Searchiness=1).



(a) nanoHUB



(b) nanoHUB



(c) nanoHUB

Figure 1: a) Visual representation of a multidimensional space in two dimensions. b) a prototypical searcher. c) a prototypical wildcatter.

In this paper we show the analysis of a whole user population using a specific tool and fuse that data set with specific classroom users.

III Classroom Clusters

To demonstrate our ability to fuse different data sets from our datastore we pick two different class clusters with significantly different characteristics as depicted in Figure 2. Class C12 is a class that reoccurred in 15 times between 2008 and 2018. We have *pntoy* simulation data from 7 classes within 2014 to 2018 for 109 users who ran 180 sessions. Historically, we do not have the simulation data from all users in that time frame. Going forward in the future we have developed a simulation caching system where all Rappture simulations are stored and users will receive stored solutions if they exist. The cluster view in Figure 2 shows a subset, the individual class held in the fall 2015 with 40 students who ran 80 sessions. C12 only uses *pntoy*. Class C16 uses 6 different tools throughout a semester. *pntoy* is one of these 6 tools used by 20 users in 29 sessions. The visual cluster representation in Figure 2 clearly shows the temporal behavior of 7 users who have used all 6 tools in the class. In the next section we will compare the behavior in these classes against all available data and against all self-study users within the same region (Texas).

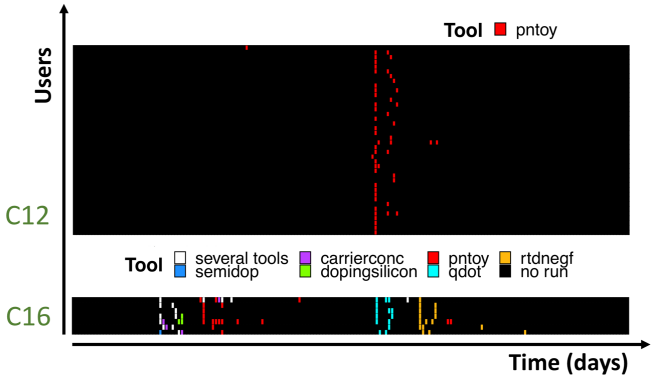


Figure 2: Visual representation of two temporal usage patterns in two different classes. The horizontal axis represents time in units of days. The vertical axis stacks different users within a cluster. 40 users in C12 of fall 2015 use *pntoy*. C12 only uses *pntoy*. Class C16 uses 6 different tools throughout a semester. *pntoy* is one of these tools used by 20 users. 7 users utilize all 6 tools as depicted.

IV User Behavior Distributions

Figure 3 shows the Searchiness distribution of all 2,747 geo-located simulation sessions of *pntoy* by 1,865 users in the time frame of 2014-2018. The complete distribution of all runs shows clear peaks around 0 (wildcatters) and 1 (searchers).

Class cluster C12 is a subset of all the available data consistent of 40 users with 80 sessions. This cluster usage uses only a single tool in the whole class. Wildcatter behavior appears to dominate this class C12. In contrast the smaller class that uses in total 6 tools, including *pntoy* with 20 users and 28 sessions shows a distribution that seems to indicate more searchers.

Finally, we look at a third population within the users. These are all the geo-located users in Texas (the location of C12 and C16) who have not been identified as participants in

in any classes. We title this group of 20 who ran 36 simulation sessions in *pntoy* as "self-study" users. These users show yet a different distribution of Searchiness compared to the other populations.

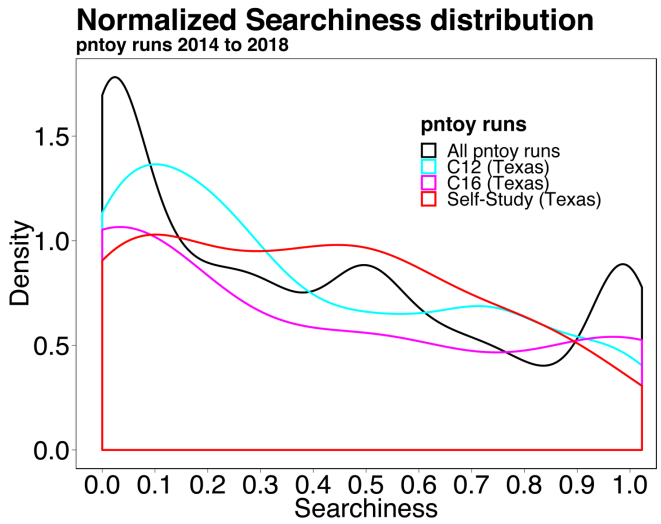


Figure 3: Normalized Searchiness density for four nanoHUB populations. All *pntoy* runs contain 1,865 users with 2,745 sessions. C12 Texas contains simulation data from 109 users running *pntoy* in 180 sessions in 7 classes from 2014-2018 (we do not have the simulation data of all users in those classes). C16 is a class that occurred once in the spring of 2015 and uses 6 tools. 20 users utilized 28 sessions. The Self-Study users populations are all 20 geo-located users in Texas who ran 36 sessions who were not associated with a formal class in the time fram of 2014-2018.

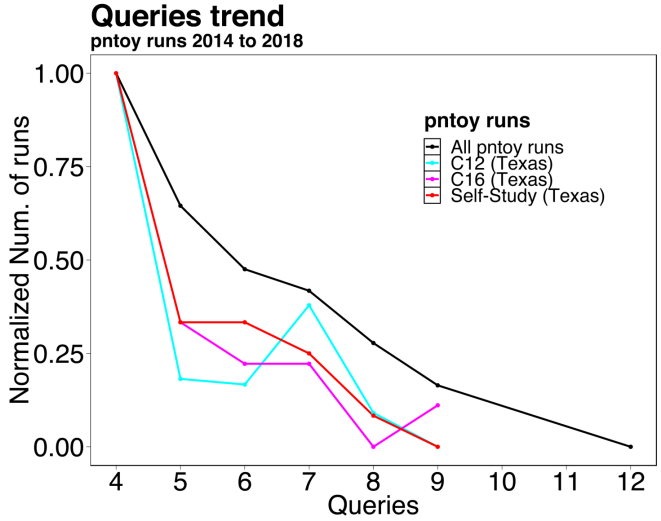


Figure 4: Normalized number of queries for four nanoHUB user populations described in Figure 3.

Next to Searchiness, which is a computed model metric, we can also look at a simple raw number, which is the number of queries each individual has performed within a single tool session. Within each tool session a user can execute the tool multiple times and compare results as visualized in Figure 1. Figure 4 shows the normalized distribution of queries executed by the 4 different populations we examined in Figure 3. The number of queries does not reveal much information except that the overall population runs more queries than the 2 Texas classes and the Texas self-study users. The classes and self-study users show a rather strong drop-off for more than the minimal queries of 4, which is the minimal number of queries needed to define Searchiness. Initial analysis does not seem to indicate a strong correlation to between Searchiness and number of queries.

VI Conclusion

We report the development of a nanoHUB infrastructure that begins to enable the study of user behavior in virtual worlds of simulation tools. We use the previously published model index Searchiness and compute it for a complete data set of simulation sessions within a specific tool. We fuse data sets of class cluster identification with the model index Searchiness and number of queries. No surprising results are seen or critical insight gained at this stage. We observe in the data that different user populations appear to behave differently in terms of Searchiness and classes seem to appear similar in terms of number of queries. At this stage the data opens new vectors for questions such as:

- Do all single-tool classes have similar behavior?
- Do classes with more diverse tool use or better scaffolding foster more search-like behavior?
- Can similar behavior differences be seen with the other tools that are used in classes?
- What does a peak in Searchiness value of 0.5 mean? Do we need to refine the Searchiness index?
- Do we need to identify other behavioral metrics in addition to Searchiness?
- Do the users who use other nanoHUB material outside the tools behave differently than the ones that use tools only?

We conclude that this work is a first demonstrator that indicates that we can assess the simulation behavior of different user populations inside nanoHUB. We plan to refine these metrics and classifiers to gain more insights on the user behavior, and ultimately influence their behavior during use.

Acknowledgements

Funding by the US National Science Foundation under Grant Nos. EEC-0228390, EEC-0634750, OCI-0438246, OCI-0721680, and EEC-1227110 as well as Purdue University is gratefully acknowledged.

References

[1] N.H. Kapadia, J.A.B. Fortes, M.S. Lundstrom, The Semiconductor Simulation Hub: A network-based microelectronics simulation laboratory, Proceedings of the Twelfth Biennial Conference:

University/Government/Industry Microelectronics Symposium, 1997, IEEE Xplore, DOI: 10.1109/UGIM.1997.616686

[2] Michael McLennan (2005), "Add Rappture to Your Software Development - Learning Module," https://nanohub.org/resources/240.

[3] Michael Mclennan, Rick Kennell, HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering, IEEE Computing in Science and Engineering 12(2):48 – 53, 2010 DOI: 10.1109/MCSE.2010.41

[4] Wei Qiao, Michael McLennan, Rick Kennel, David Ebert, Gerhard Klimeck, "Hub-based Simulation and Graphics Hardware Accelerated Visualization for Nanotechnology Applications". IEEE Transactions on Visualization and Computer Graphics, Vol. 12, Issue: 5, Page(s): 1061-1068, Sept.-Oct. 2006;doi : 10.1109/TVCG.2006.150

[5] Gerhard Klimeck, Michael McLennan, Sean Brophy, George Adams III., Mark Lundstrom, "nanoHUB.org: Advancing Education and Research in Nanotechnology", IEEE Computers in Engineering and Science (CISE), Vol. 10, Issue: 5, Page(s): 17 - 23, Sept.-Oct. 2008;doi:10.1109/MCSE.2008.120

[6] Lynn Zentner, Steven Clark, Krishna Madhavan, Swaroop Shivarajapura, Victoria Farnsworth, Gerhard Klimeck, "Automated Grid-Probe System to Improve End-To-End Grid Reliability for a Science Gateway", Proceedings of TeraGrid 2011 conference. July 18-21, 2011, Salt Lake City, ACM proceedings, ISBN: 978-1-4503-0888-5;doi:10.1145/2016741.2016789

[7] Michael Zentner, Nathan Denny, Krishna Madhavan, Swaroop Samek, George Adams III., Gerhard Klimeck, "Using Automatic Detection and Characterization to Measure Educational Impact of nanoHUB", Proceedings of the 13th Gateway Computing Environments Conference, September 25-27, 2018, Austin, TX

[8] Krishna Madhavan, Michael Zentner, Gerhard Klimeck, "Learning and research in the cloud", Nature Nanotechnology 8, 786–789 (2013); doi:10.1038/nnano.2013.231

[9] Dragica Vasileska, Matteo Mannino, Michael McLennan, Xufeng Wang, Gerhard Klimeck, Saumitra Raj Mehrotra, Benjamin P Haley (2014), "PN Junction Lab," https://nanohub.org/resources/pntoy. (DOI: 10.21981/D3GH9B95N).

[10] https://nanohub.org/usage/tools provides nanoHUB tool listings ranked by various criteria, such as number of users, number of simulations, wall clock time, etc.

[11] Mark R. Pinto, Kent Smith, Muhammad Alam, Steven Clark, Xufeng Wang, Gerhard Klimeck, Dragica Vasileska (2014), "Padre," https://nanohub.org/resources/padre. (DOI: 10.21981/D30C4SK7Z).

[12] Nathan Denny, Gerhard Klimeck, Michael Zentner, "Visualizing User Interactions with Simulation Tool", Proceedings of the 13th Gateway Computing Environments Conference, September 25-27, 2018, Austin, TX