# Report of Research: Engineering a Software Solution for finding the WSB Triple

This paper outlines a collaborative research project between the University of Mary Washington, UMW, and the Dahlgren Naval Surface Warfare Center, DNSWC. The research project was conducted by the following people:

**Dr. Jeff Solka, DNSWC**
**Dr. Allen Parks, DNSWC**
**Kristin Ash, DNSWC**

**Dr. Melody Denhere, UMW (Supervising Faculty)**
**William Etcho, UMW (student researcher, Math)**
**Josiah Neuberger, UMW (student researcher, Computer Science)**

The following faculty provided consultation when needed:

**Dr. Debra Hydorn, UMW (Math)**
**Dr. Stephen Davies, UMW (Computer Science)**

---

The following resources were vital to this research project:

***Quantifying Long-Term Scientific Impact by Wang, Song, and Barabási:***
http://www.sciencemag.org/content/342/6154/127.full.html
http://www.sciencemag.org/content/suppl/2013/10/02/342.6154.127.DC1.html

**Dr. Allen Parks was consulted in the following key areas:**
*The general math to compute the Wang-Song-Barabási Triple*
*Understanding the citation prediction mathematics and examining their uncertainties*

**Dr. Jeff Solka was consulted in the following key areas:**
*Specific constraints to consider when implementing Dr. Parks general math solution to the WSB Triple*

**C.S. Kristin Ash was consulted on various aspects of the project to include:**
*Implementation details, software design, debugging, and data mining/interpretation*

This paper is into the following sections:



# I. Introduction and Background

The purpose of this paper is to outline the various aspects and background information necessary to implement a software system that computes the WSB Triple given a paper's citation history of at least 5 years or more. The software system described herein uses a numerical Newton-Raphson algorithm to converge on the WSB Triple based on the general math developed by Dr. Allen Parks. The WSB Triple can then be used as input for some prediction formulas extracted from Wang, Song, and Barabási's paper to quantify future impact of the discovery.

In the paper *Quantifying Long-Term Scientific Impact by Wang, Song, and Barabási, WSB,* a paper's previous citation history was shown to provide important prediction of future citations using a series of formulas, variables, and constraints derived in their research and outlined in the paper. These three variables were dubbed the WSB Triple by Dr. Parks.[1] [2]

1. $\lambda$ - Relative Fitness
2. $\mu$ - Immediacy
3. $\sigma$ - Longevity

In order to understand the WSB Triple, we must briefly look at the various aspects controlling a paper's likelihood of increasing its citations. Dr. Wang et al identifies three mechanisms in their paper that causes citation change. First, preferential attachment explains the notion that once existing papers receive citations they are more likely to receive additional citations in the future based on this increased visibility. This concept of preferential attachment is the beginning of justification for predicting a paper's future citations from its

[1] Parks, Allen Dr. (Private Communication: Computing the Wang-Song-Barabási Triple)

previous citation history. Second, as a paper ages its ability to obtain new citations diminishes. This mechanism is best described by thinking about how a paper outlining the discovery of wheels is going to lose new citations to papers about motorized vehicles. At first, the wheel's paper will be cited by the motorized vehicle research, but eventually it will be integrated into these new papers. This concept of a paper aging is captured in the WSB triple by $\mu$, which tracks the time it takes a paper to reach the maximum number of citations and $\sigma$, which tracks the decay rate using a log-normal survival probability ("fading novelty").[2]

The final mechanism, fitness, is not so well defined by WSB. The basic idea of fitness is that a paper will have a degree of "novelty" and "importance of discovery" associated with it that will help garner new citations. However, WSB rightly points out that these two concepts are hard to generalize to measurable values across papers. They instead measure a paper's fitness by looking at the overall response to a paper by the community, which according to them is a measurable quantity they term as "Relative Fitness".[2]

In the next section, we will continue discussing the WSB paper by considering a few important prediction formulas that WSB derived during their research. Moving forward from there, we will discuss the software solutions used to engineer the Newton-Raphson Convergence Search algorithm for solving the WSB Triple. Important variables and formulas will be covered alongside major design decisions. The pseudo-code for the algorithm will be provided and discussed. We will also consider our success and failures at reproducing certain values discussed in the WSB paper with our algorithm. Finally, this paper will wrap up with a discussion of the results from running the NRC Search on several hundred PubMed papers.

## II. Using the WSB Triple to Quantify Impact

The following formula from the WSB paper can be used to graph the WSB solution to see the fit against the plotted citation history:[3]

$$c_y = m\left[e^{\lambda * pnorm\left(\frac{\ln t - \mu}{\sigma}\right)} - 1\right] \qquad (1)\ [4]$$

Remember the WSB Solution/Triple is the $(\lambda, \mu, \sigma)$ pairing acquired from the citation history. Equation #1 gives the cumulative citations for a paper at a time 't'. If you graph 'c' on the y-axis and 't' on the x-axis on top of a plot of the citation history then you should be a able to clearly visualize the fit of the WSB solution. It follows that one can project this formula into the future to predict future citations.[5] In section #10 of this paper, we will show that we reproduced these graphs using our engineered software system within a margin of error.

[2] Wang, Song, Barabási [*Quantifying Long-Term Scientific Impact*] (Science 342, 127 (2013) pg. 128 -129)
[3] Wang, Song, Barabási [*Quantifying Long-Term Scientific Impact*] (Science 342, 127 (2013) pg. 129 equation 3 and 4)
[4] Refer to WSB paper pg. 128: Fig. 1 Graph E for an example.
[5] Refer to WSB pg. 131: Fig. 4 Graph A and B for examples.

An obvious question comes to fruition: what is the maximum number of citations a paper will receive in its lifetime? WSB answers this question by considering the above formula as $t \to \infty$:

$$c_{max} = m\left(e^{\lambda} - 1\right) \qquad (2)\ [2]$$

Taking the limit as $t \to \infty$ we obtain the ultimate impact, which represents the total number of citations a paper will garnish in its lifetime and is shown to be independent of μ or σ.

## III. Assumptions and Uncertainties

- The value 'm' is a constraint that is loosely defined by WSB to be a proxy of "initial attractiveness" of a paper. Basically, we can calculate 'm' for a group of similar papers, such as the PubMed journal, by calculating the average number of references contained in each new paper. However, WSB points out that this number is best fixed for all papers at 'm=30' to better compare different papers. Furthermore, they constrain 'm' to a value close to the expected value of 'c', which makes this value difficult to generalize without some significant study.[6] The paper states that 'm' is a global parameter and that our results do not depend on our choice of 'm'.[2] We found that within a certain range 'm' had little effect on our solutions, but if we used extreme values such as m=2.9 or m=100 there were cases in which our predictions were skewed. In the end, we used their recommend value of m=30 for our PubMed results.

- We obtained a few '–λ' solutions which produced a graph with a negative citation pattern. As this is not possible we chose to disregard those solutions and instead use the next positive root.

- Our algorithm searches for roots for all values of μ and σ from 0.1 to 10. In each case our method only found one root in that range and we assumed that was a unique solution.

## IV. Engineering a Software System to Solve WSB

We chose to prototype our software system on Windows using the R programming language. R includes the key functionality needed to get started: CDF/PDF functions, plotting, and reading in data from CSV files. On the downside, R is not as user friendly as Java in many ways such as documentation of the API. In the end, we coded our software system in both programming languages and will cover important design decisions for our Java and R code.

R provides a built in normal distribution function and a normal probability density function: 'pnorm' and 'dnorm'. R provides support for plotting and saving graphs through the 'plot' function and the dev.copy/dev.off

[6] Wang, Song, Barabási (Science 342, 127 (2013) Supplement Material)

functions (saving graphs).[7] In order to keep track of the many variable values needed to solve for the WSB solution, one needs to be familiar with basic data structures in R: lists, data frames, matrices, vectors and basic misc. functions: setwd, getwd, read.table c(), rbind, cbind and the element operator: '|' (accessing elements inside lists).[8]

We used several open source libraries to support our Java implementation. We used the Apache Commons library, which is licensed under the Apache License, Version 2.0.[9] In order to get 'pnorm' and 'dnorm', we used the Apache.Commons.Math.Distribution package to write these functions:

```
nd: NormalDistribution
pnorm:double (x: double) { return nd.cumulativeProbability(x); // normal cdf }
dnorm:double (x: double) { return nd.density(x); //normal pdf }
```

In order to get a matrix nature in Java, we used the Apache.Commons.Math.Linear subpackages: RealMatrix, LUDecomposition, DecompositionSolver, and MatrixUtils. In order to plot the data and graph the WSB curve, we used the Java library jmathplot.[10]

We organized the paper data in CSV files with each paper's data in a single row. The first two columns provide identifying information: some kind of integer id and a 4 digit year. The remaining columns are dedicated to the citation history of the paper. Each one should contain a year's worth of citations. If the paper received no citations for a year then the file should contain 0 in that column:

3040403,1950,3,4,0,10,0,0.....,0

This paper received 3 citations from time=0 (publishing) to time=1 year, 4 in the second, 0 in the third, 10 in the 4th, etc. The NRC Search requires the data to be arranged in a two column matrix translated into mostly uniform day-based citation pattern:[11]

| Timestamp, | Citation # | Timestamp, | Citation # |
|---|---|---|---|
| $t_1$=121.66, | 1 | $t_5$=547.50, | 5 |
| $t_2$=243.66, | 2 | $t_6$=638.75, | 6 |
| $t_3$=365.00, | 3 | $t_7$=730.00, | 7 |
| $t_4$=456.25, | 4 | $t_8$=... | 8 |

[7] Geyer, Charles Dr.[Probability Distributions in R] Link: http://www.stat.umn.edu/geyer/old/5101/rlook.html
Spector, Phil [Saving Plots in R] Link: http://www.stat.berkeley.edu/classes/s133/saving.html
[8] Rodríguez, Germán [Introducing R] Link: http://data.princeton.edu/R/gettingStarted.html
Rodríguez, Germán [Reading and Examining Data] Link: http://data.princeton.edu/R/readingData.html
[9] The Apache Commons Library can be found at http://commons.apache.org
[10] The jmathplot library can be found at http://code.google.com/p/jmathplot/
[11] Solka, Jeff Dr. (Private Communication: Implementation specifics for the ADP\ Solution of the WSB Equations)

The WSB supplemental material which outlines the WSB solution $(\lambda, \mu, \sigma)$ can be obtained by solving equation (S22) numerically. In order to do this, (S22) was optimized with respect to each WSB parameter producing a system of three nonlinear equations. In private communication with Dr. Parks, $\lambda$ was eliminated resulting in a system of two nonlinear functions found below in equations 3 and 4.

$$\hat{l} = \ln\lambda + \langle \ln\left(\frac{i-1}{N}\right) + \hat{m}\rangle + \langle \ln P_i + \lambda\Phi_i\rangle - \lambda(1+\hat{m})\Phi_T + \ln N \qquad \text{(S22)}$$

$$f(\mu,\sigma) = [(1+\hat{m})pnorm(x_t) - \langle pnorm(x_i)\rangle]\langle x_i\rangle - \langle dnorm(x_i)\rangle + (1+\hat{m})dnorm(x_t) \qquad (3)$$

$$g(\mu,\sigma) = [(1+\hat{m})pnorm(x_t) - \langle x_i\rangle](\langle {x_i}^2\rangle - 1) - \langle x_i * dnorm(x_i)\rangle + (1+\hat{m}) * x_t * dnorm(x_t) \qquad (4)$$

The Newton-Raphson method can be used to solve these two equations numerically producing a $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ solution which can be used to find lambda through equation #5 below:

$$\lambda = [(1+\hat{m})pnorm(x_t) - \langle pnorm(x_i)\rangle]^{-1} \qquad (5)$$

The Newton-Raphson method solves an equation by finding an approximation of a zero for the function. The unknown variable values can then be extracted from this approximate root. The basic idea of the Newton-Raphson method is to:

1. Pick a point, '$p_1$', on the function line.
2. Graph the tangent line at that point $(p_1, f(p_1))$.
3. Find the x-intercept of this tangent line to use as a new point closer to the desired root.
4. Use this $(p_2, f(p_2))$ can be used to graph a tangent line resulting in $p_3$.

This iteration pattern is repeated until $f(p_n) = 0$ where we finally arrive at the approximated zero of the function. The tangent lines described will have slopes of $f'(p_1), f'(p_2), \ldots, f'(p_n)$. Using this derivative of the series of points and the knowledge that '$p_n$' is the x-intercept of the previous iteration's tangent line we can arrive at the generalized formula for the Newton-Raphson method:[12]

$$p_{n+1} = p_n - f'(p_n)^{-1} f(p_n) \qquad (6)$$

On the next page, we discuss how to expand the Newton-Raphson for a system of non-linear equations.

[12] Ellis, R and Gulick, D (2003) *Calculus* [Sixth Edition]. Mason, Ohio: Thomson Learning Custom Publishing

The Newton-Raphson method can be expanded to approximate a root for a system of non-linear equations by using a matrix equation along with an inverse of a Jacobian matrix made up of the partial derivatives of the system. Dr. Parks provided the two non-linear functions #3 and #4 on the previous page. A Jacobian matrix can be built by calculating the partial derivatives of 'f' with respect to mu and then sigma. Repeating this process with function 'g' yields the following Jacobian matrix:

$$j = \begin{pmatrix} \frac{\partial f}{\partial \mu}\Big|_n & \frac{\partial f}{\partial \sigma}\Big|_n \\ \frac{\partial g}{\partial \mu}\Big|_n & \frac{\partial g}{\partial \sigma}\Big|_n \end{pmatrix} \quad (7)$$

$$x_{n+1} = x_n - j^{-1} y_n \quad (8)$$

$$x_n = \begin{bmatrix} \mu_n \\ \sigma_n \end{bmatrix} \quad (9)$$

$$y_n = \begin{bmatrix} f_n \\ g_n \end{bmatrix} \quad (10)$$

Since we are taking the inverse of the Jacobian matrix we must first check that it is non-singular. Each time we solve the matrix equation we produce a $x_{n+1} = \begin{bmatrix} \mu_{n+1} \\ \sigma_{n+1} \end{bmatrix}$ that is hopefully converging towards the desired root. If our initial guess for (mu, sigma) is bad the Newton-Raphson method might be going away from the root and fail to converge at all. If multiple roots exist the Newton-Raphson method could converge on the wrong root. The WSB material really does not clarify the constraints or method for finding the right root; however, we discuss later in the PubMed results that our system finds an acceptable root in most cases.

A few minor details must be considered before moving onto the algorithm for the Newton-Raphson method. The CDF and PDF functions are using the following calculations:

$$x_T = \frac{\ln T - \mu}{\sigma} \quad , \quad x_i = \frac{\ln t_i - \mu}{\sigma}$$ :: (11) where T is the last time stamp in the citation history.

The notation $\langle x \rangle$ is defined as the 'expected value of x' and can be calculated by:

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x$$ :: (12) where N is the total number of citations for a paper in time [ 0, T ].

$$\hat{m} = {}^{m}/_{N}$$ :: (13) where m is the constant defined earlier to be a suggested value of 30.

Our Newton-Raphson method leaves the calculation of these values and the partials to a separate function, which takes each $(\mu, \sigma)$ as input. These values are stored in a list structure called 'l'.

**The pseudo code of our Newton-Raphson Numeric Solver can be found on the next page.**

**Newton-Raphson Solver: (lamda, mu, sigma)**

```
    requires   : data:matrix[x rows, 2 cols] - (day timestamp, cumulative citation)
               : mu guess, sigma guess
               : m - average number of references for each new paper
               : t - final time-stamp from data matrix
               : n - cumulative citations
               : l:list - structure to track iteration data (null to start)
               : iteration - (0 to start)
               : tolerance - Convergence Tolerance(.1 to start)

if (iteration > max_iteration) then //Newton-Raphson is not converging
    return 'no solution' ie: (null, null, null)
else if (tolerance < 1e^-8) then
    solve for lambda using equation e(lambda) with (mu,sigma)
    return 'solution' ie: (lambda, mu, sigma)
else
    //calculate values using the current iteration's (mu, sigma)
    get 'l' from iteration data: function 'getIterationData' with (mu, sigma)

    //[rows, columns] indexed from 1
    set xn:matrix[2 by 1] to (mu, sigma)
    set yn:matrix[2 by 1] to (l->fn, l->gn)

    set jacobian:matrix[2 by 2] to  set row1 to (l.df_dmu, l.df_dsigma),
                                        Set row2 to (l.dg_dmu, l.dg_dsigma)
    if (jacobian is singular) then
        set iteration to max_iteration ie: return 'no solution' on next iteration
    else
        set solution to xn-jacobian^-1(yn) //matrix equation
        set tolerance to sqrt((solution.mu - xn.mu)^2 + (solution.sigma - xn.sigma)^2)

        set l.mu to solution.mu
        set l.sigma to solution.sigma

        return recursive call with data, l.mu, l.sigma, m, l, iteration, tolerance
```

In private communication with Dr. Parks, our intuition was to find the initial guess for the Newton-Raphson by solving for linearized versions of the non-linear (S26) equations; however, the Newton-Raphson Solver is very sensitive to slight variations in the (mu, sigma) initial guess. We overcame this problem, by implementing a convergence search algorithm, which tries many initial guesses of (mu, sigma) starting at 0.1 to a specified upper parameter. The algorithm was found to be very successful at converging on a WSB Solution in all but a few of our tests using the PubMed data.

```
Newton-Raphson Convergence Search: list of (lamda, mu, sigma) solutions
    requires: data:matrix[x rows, 2 cols] - (day timestamp, cumulative citation)
                : start - starting point (default=0.1)
                : upperLimit(mu, sigma) - (default=(10,10))
                : step - how fast to approach upper limit (default=1)
                : m - average number of references for each new paper
                : wasAlreadyRun - (default=false)

initialize matrix:list
initialize solutions:list
initialize lambdas:list

for (set mu0=’start’ until ‘upperLimit.mu’ by ‘step’)
    for (set sigma0=’start’ until ‘upperLimit.sigma’ by ‘step’)
        get ‘answer’ from newtonRaphsonSolver with data, mu0, sigma0, m

        if (answer.lambda is not null) then
            set row to answer.(lambda, mu, sigma)
            set isUnique to true

            foreach (l in lambdas)
                if ( (answer.lambda is negative) or ( ‘l’ -answer.lambda meets tol. check of 1e-2)
            )   then
                    set isUnique to false
                    break //exit foreach loop because solution is not unique.
            //end foreach

            if (isUnique) then add to ‘solutions’: answer.(lambda, mu, sigma)

            add to ‘lambdas’: answer.lambda
    //end inner for loop
//end outer for loop

if (not:wasAlreadyRun and solutions is empty) then
    return recursive call with data, start, upperLimit(mu,sigma), step=.1, m, wasAlreadyRun=true

else return ‘solutions’ list
```

## V. Reproducing key WSB Results

The WSB paper provided one WSB solution along with its citation history, which we used to verify our algorithm.[13] The WSB solution for this test data from the paper is $(\lambda = 2.87,\ \mu = 7.38,\ \sigma = 1.2)$. Our NRC Search algorithm found the WSB solution to be $(\lambda = 2.78,\ \mu = 7.53,\ \sigma = 1.02)$ based off of extracted citation history using a digitizer on the WSB test data. Our solution was within (3.14%, 2.03%, 15.00%) of each of the WSB values using equation #14 below on $\lambda, \mu, and\ then\ \sigma$ for finding the percentage difference. Figure #1 below shows our WSB curve (red line) graphed against the cumulative citation history (blue line). We also verified that our curve continued to represent the citation history when using only 5 years (yellow) and ten years (green) of training to find the WSB solutions. Reproducing this result was the first step to verifying that our algorithm works correctly.

$$\frac{|x - x_{wsbpaper}|}{x_{wsbpaper}} \qquad (14)$$

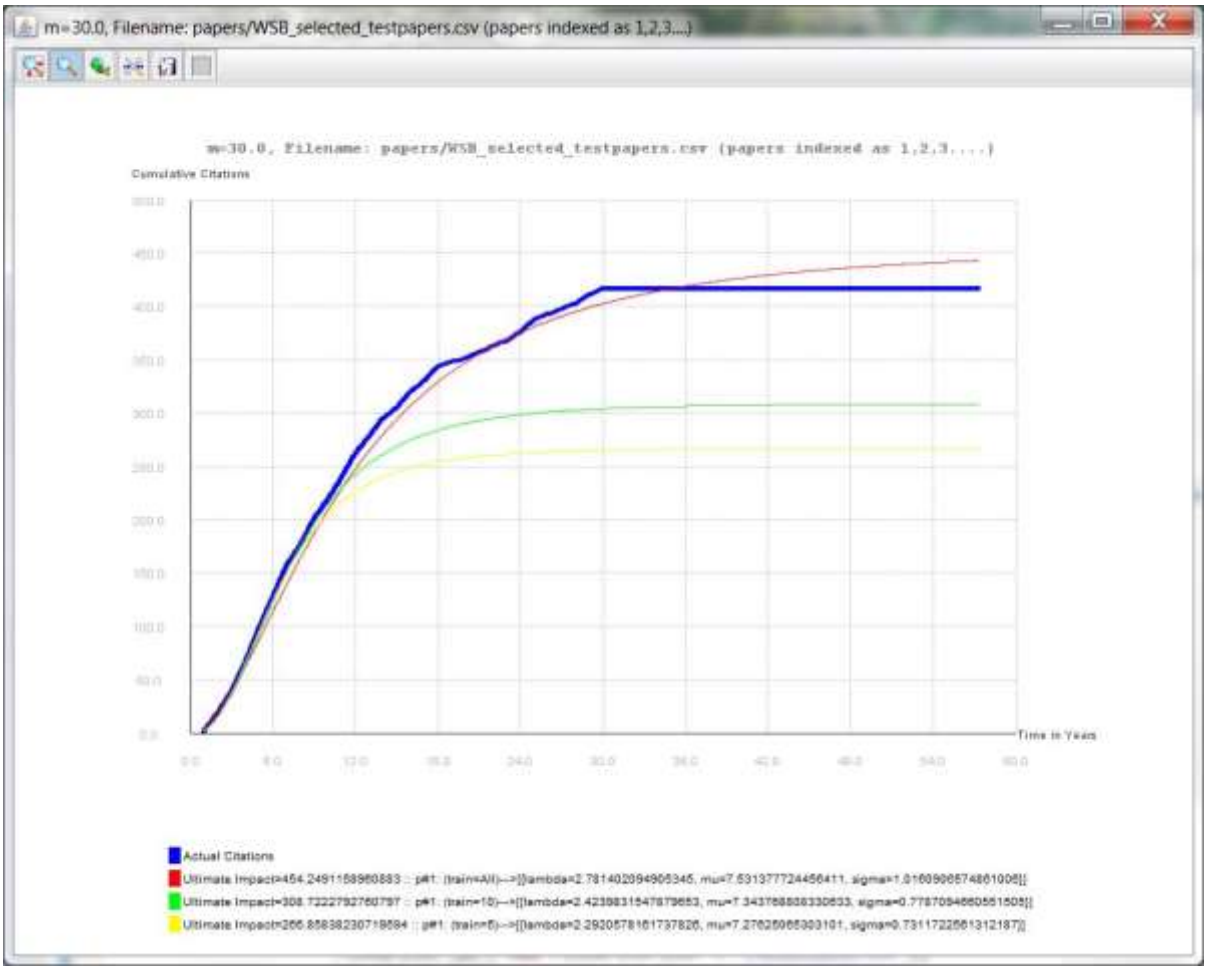


Figure #1: Reproduction of key WSB result #1

[13] Refer to page 128 figure #1 graph 'E' of the WSB paper.

The WSB paper provided three more graphs (no WSB values provided for these), which we were able to reproduce in figure #2 below. The citation history for these graphs was extracted using a digitizer on the WSB graphs.[14] Our curves did not match the citation history as closely as their graphs show is possible; however, the initial citation history from year=0 to year=5 had so many citation points that the digitizer could have introduced significant error by this factor alone.

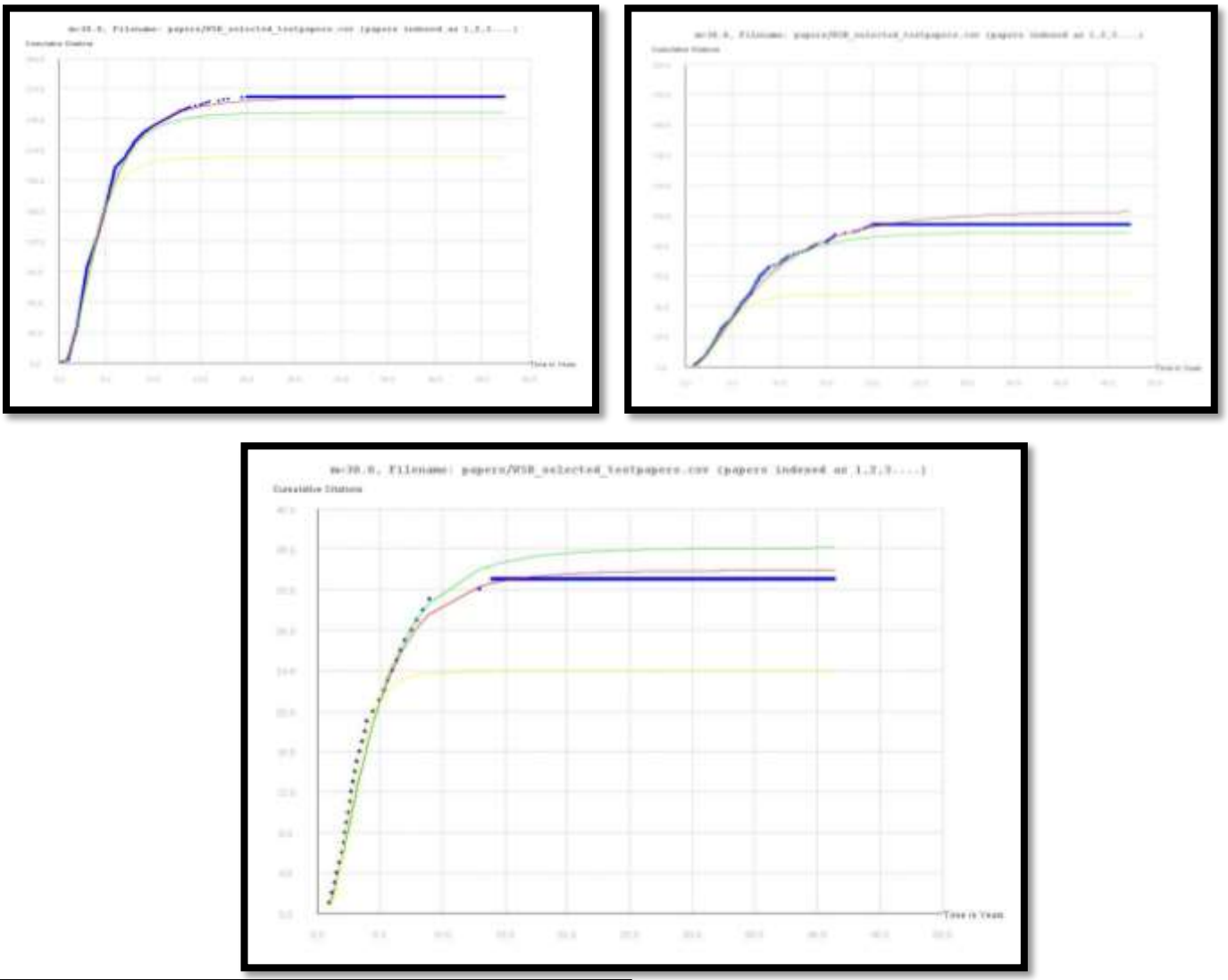

Figure #2: Reproduction of three key WSB results.

Reproducing these four curves along with the WSB solution verified that our algorithm was producing desired results within the limited framework of the original WSB test cases. In the next section, we expand our research to the real world using our algorithm to predict citations for different papers.

[14] Refer to WSB paper's graph 'A' (5 years of training) and 'B' (10 years of training) from page 131 figure #4.

## VI. Verification and Application: PubMed Results

We wanted to further test the effectiveness of our numerical approach for solving for a WSB solution so we tested our algorithm using papers from a different source. We selected papers from the PubMed database, which tracks citations for biomedical literature for the US National Library of Medicine and the National Institutes of Health. We continued to use the same constraint used with the above papers: only using papers which received at least 10 citations in the first 5 years. We relied on the same approach above to verify our results: comparing the WSB curves produced when using 5 (yellow), 10 (green), and all (red) years of data for training. Finally, we determined the quality of our WSB solution by examining the ultimate impact (total predicted citations from our algorithm) and comparing it with the cumulative total citations from the citation data. We tested our algorithm on 620 papers published in the 1950s to 1980s finding WSB solutions for almost all of them. We have selected a sample of graph types repeatedly seen in our WSB solutions to discuss below:

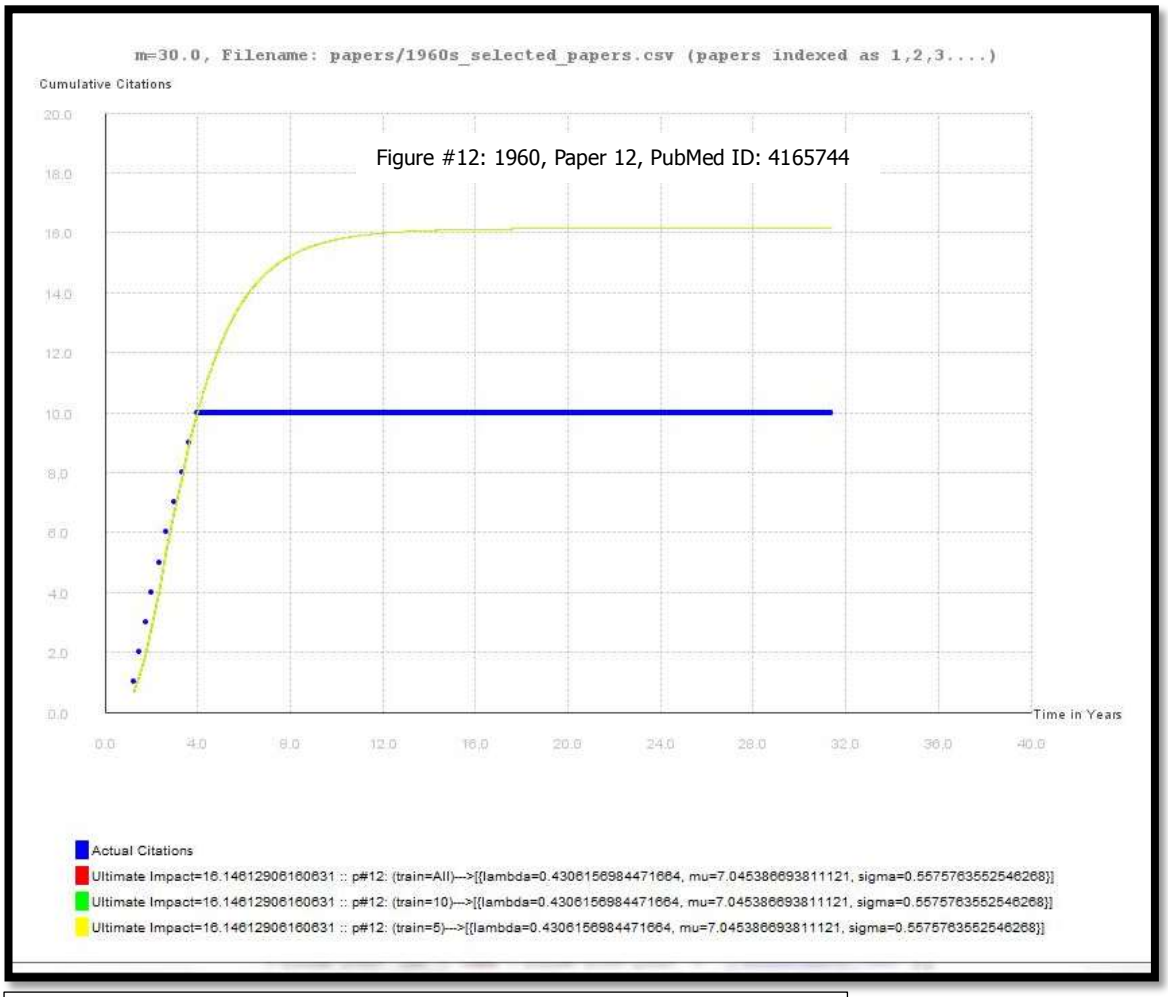


Figure #5: WSB graph with same solution for all three training periods.

In figure #5 on the previous page, our NRC Search found a WSB solution when using 5, 10, and then all years of citation history as training. Each solution converged on the same WSB values, which predicted an ultimate impact that was 65% over predicted when compared to the actual citation data (using equation #15 below for finding the percentage difference). We see this citation pattern in many of the papers, which just barely meets our criteria of at least 10 citations in the first five years, followed by almost no additional citations. Our NRC Search found many similar quality solutions to other graphs with this citation pattern. The next step to better understand these graphs would be to analyze if the predicted results improve for similar papers that have a better initial citation pattern of say maybe at least 20 citations in the first 10 years.

$$\frac{u-c}{c} \qquad (15),$$

Where 'u' is the predicted ultimate impact (maximum possible citations) and 'c' is the cumulative citations calculated directly from the citation history. A negative indicates under prediction and no sign indicates over prediction.

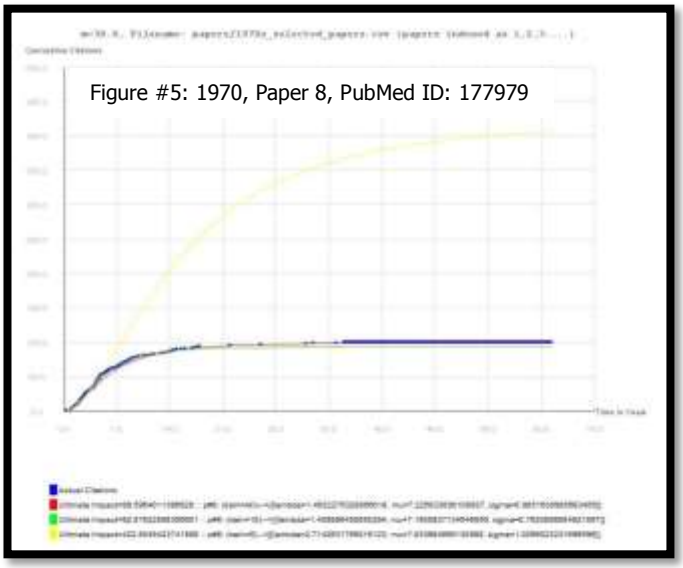

Figure #5: 1970, Paper 8, PubMed ID: 177979

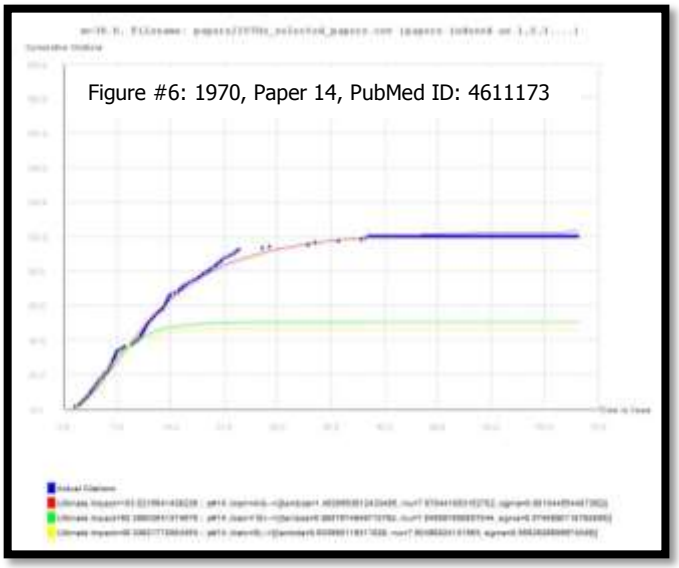

Figure #6: 1970, Paper 14, PubMed ID: 4611173

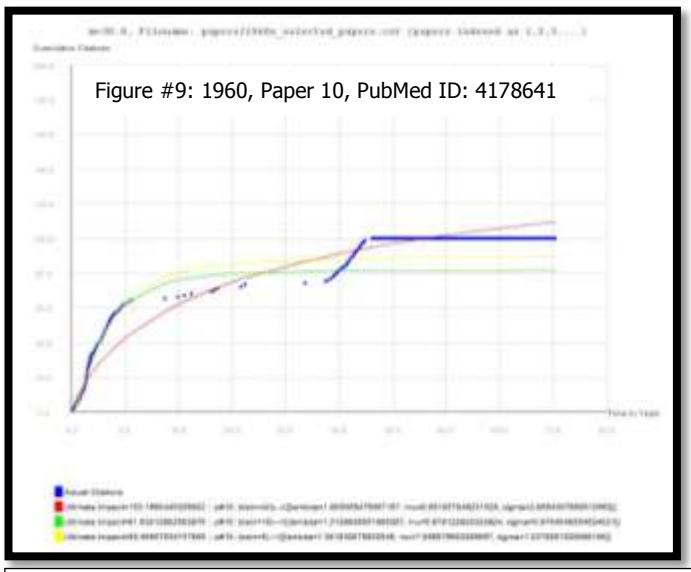

Figure #9: 1960, Paper 10, PubMed ID: 4178641

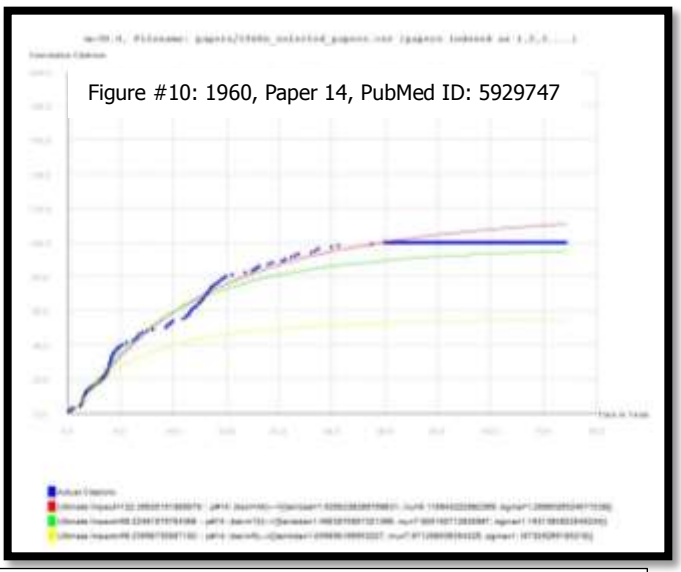

Figure #10: 1960, Paper 14, PubMed ID: 5929747

Four paper's WSB solutions using (5, 10, all) training periods graphed against the citation history.

In the above four papers, the curves show the second kind of pattern we observed in many of the graphs. Three unique WSB solutions were found with increased training providing better prediction accuracy. For instance in figure #5, the yellow curve indicates using 'training=5', producing an ultimate impact that over predicted by 322%; however, increasing the training to 10 years corrected the prediction to under estimating by only 7%. In figure #10, increasing training from 5 to 10 years, corrected a -44% prediction to one that was accurate to within 1% of the citation history. Figure #9 shows a solution were increased training decreased the prediction quality for the ultimate impact. However, we've included it here because of the interesting spike of new citations at around 30 years, which still did cause the solution to be an extreme: 10 years of training provided an ultimate impact within -18% of the citation history.

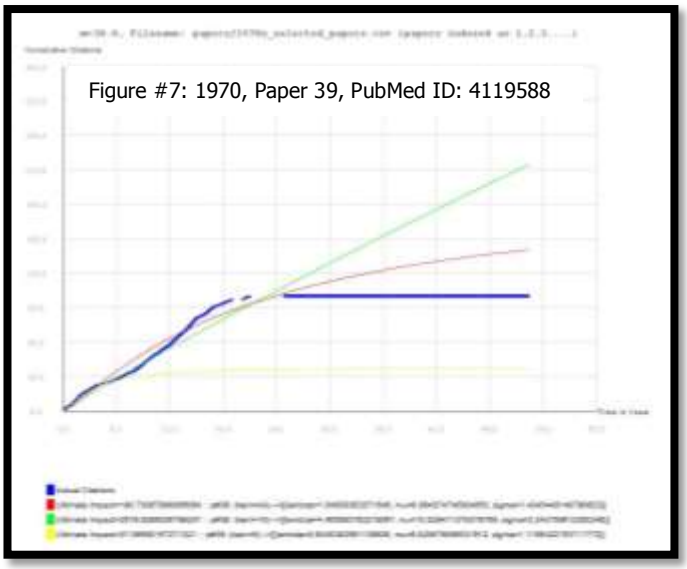

Figure #7: 1970, Paper 39, PubMed ID: 4119588

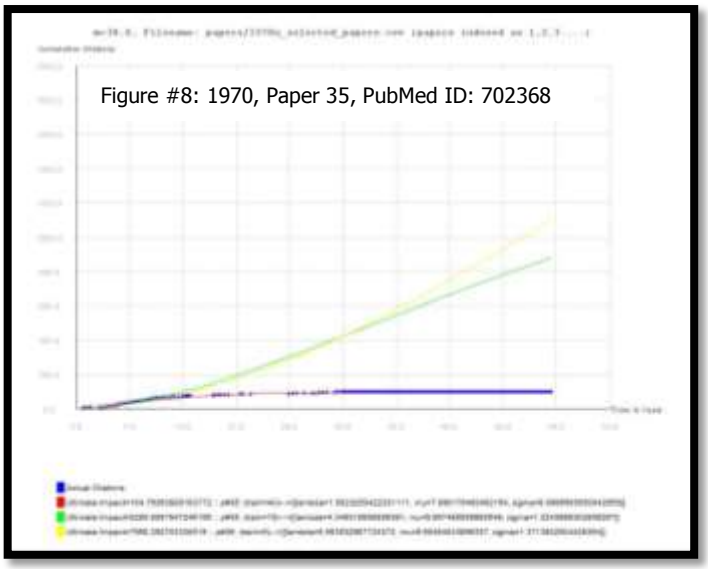

Figure #8: 1970, Paper 35, PubMed ID: 702368

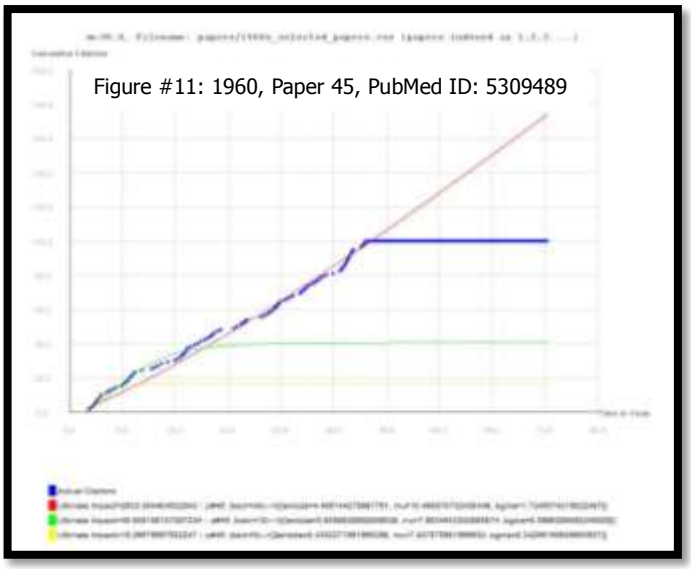

Figure #11: 1960, Paper 45, PubMed ID: 5309489

Three paper's WSB solutions using (5, 10, all) training periods graphed against the citation history.

In the above three graphs, our NRC Search found solutions that are not as useful as before. In figure #7, the 10 years of training produced an ultimate impact prediction that was 2479% off the citation data; however, the 5 year and all year WSB solutions were not so drastic. Moving to figure #8 we see similar results, which indicate

citations in the thousands except for when using all the data for training (within 8% of the actual citation history). In the last graph (figure #11), the 5 and 10 year training solutions under predicted but were still good results. The red line which uses all data for training indicates an ultimate impact in the thousands. We included these graphs to indicate and represent some of the limitations of the NRC Search algorithm. The WSB paper indicates some limitation may exists specifically stating that "citations bumps" like what we saw in figure #9 could potential cause bad predications.

## VII. Conclusion of Research Results

Our Newton-Raphson Convergence Search algorithm reproduced four key results outlined in *Quantifying Long-Term Scientific Impact* by WSB regarding prediction power of this model. Using selected papers from PubMed, our NRC Search was able to find WSB solutions that predicted citations with a reasonable degree of accuracy for many of the papers. Of the 188 papers from the 1970's when using 10 or more years for training, our NRC Search predicted ultimate impacts within $\pm$20% of the citation history data for those papers. Of the 65 papers from the 1960's, our algorithm found 20 more papers. These results are very promising and indicate that our algorithm has strong prediction power for selected kinds of papers. We showed some areas were the algorithm faced limiting factors or potential erroneous predictions. These areas must be further explored to define what kinds of papers are appropriate for use with our algorithm.

## VIII. Future Outlook and Application for the WSB Solution

The next step in accurately predicting future impact and citation counts lies in two uncertainty definitions extracted from the WSB Supplement martial:

$$\sigma_p^+ = \sqrt{\int_{k^*}^{\infty} \left(k_p - k_p^*\right)^2 P(k)dk} \quad \text{(S32a)}, \qquad \sigma_p^- = \sqrt{\int_{N}^{k^*} \left(k_p - k_p^*\right)^2 P(k)dk} \quad \text{(S32b)}$$

Taken together they provide a citation prediction envelope. By comparing our predicted number of citation to how far they deviate from the actual citations using test data we can determine whether the papers future citations are correctly predicted by the model. As we were limited in time for this project we were not able to create an algorithm to model this prediction envelope. However, this confidence range was used to determine the accuracy of the model from which our algorithm was developed and in verifying their results WSB states "it is rather rare to have a paper exiting the citation envelope".[15]

The continuation of this research is recommended because the tools developed have wide-reaching application potential. For instance, one such use for an algorithm that predicts the future citation history of papers is to help impact future investment decisions. By predicting future citations one can assess the long-term impact that papers will have and therefore give investors an idea of projects that will be most

[15] Wang, Song, Barabási (Science 342, 127 (2013) Supplement Material pg. 16)

beneficial to invest their time and money in. Another important motivation is discussed in the WSB paper, they mention how *"current measures of citation-based impact from IF to Hirsch index are frequently integrated in reward procedures, the assignment of research grants, awards, and even salaries and bonuses, despite their well-known lack of predictive power."* [2]